# Identification et Vérification du Scripteur dans des Documents Manuscrits

## Writer identification and verification in handwritten documents

**Ameur Benséfia, Thierry Paquet, Laurent Heutte**
Laboratoire PSI - FRE CNRS 2645, UFR des Sciences, Université de Rouen,
76821 Mont-Saint-Aignan Cedex, France
{Ameur.Bensefia, Thierry.Paquet, Laurent.Heutte}@univ-rouen.fr

Manuscrit reçu le 19 mai 2004

Résumé et mots clés

**Dans cette communication, nous appliquons un modèle de recherche d'information pour la tâche d'identification du scripteur. Les requêtes sont des images de documents qui sont tout d'abord projetées dans un espace de caractéristiques. La base de documents manuscrits est indexée selon le principe du modèle vectoriel de recherche d'information textuelle. L'approche exploite donc à la fois la représentation mixte image et textuelle spécifique d'un document manuscrit. Les documents identifiés à l'issue de cette étape font ensuite l'objet d'une analyse complémentaire pour vérifier les hypothèses émises. Nous proposons d'utiliser un critère d'information mutuelle pour vérifier que chacun des documents identifiés peut avoir été produit par le scripteur de la requête. Nous utilisons un test d'hypothèse à cet effet. L'approche est testée sur deux bases d'écritures différentes et montre une grande robustesse aux différentes écritures. L'approche semble donc très intéressante pour des applications à plus grande échelle nécessitant d'interroger des bases de documents manuscrits.**

**Documents manuscrits, identification du scripteur, vérification du scripteur, recherche d'information, information mutuelle, test d'hypothèse.**

Abstract and key words

In this communication we apply an Information Retrieval model for the writer identification task. Queries are handwreitten document images projected on a suitable feature set. The handwritten document database is indexed according to the vector space model originaly used for textual information. The approach uses both the image and textual description of handwritten documents. Identified documents are then processed by the verification stage. We use a mutual information criterion so as to verify that each identified document can have been written by the writer of the query. Decision operates using an hypothesis test. The approcah is evaluated on two different database and proves to be robust to the variability of handwriting. Perspectives are oriented towards the use of large handwritten document database

Handwritten Documents, Writer Identification, Writer Verification, Information Retrieval, Mutual Information, Hypothesis Test.

# 1. Introduction

Cet article aborde la problématique de l'analyse automatique des documents manuscrits pour l'aide à l'expertise. En effet, dans certains cas d'expertise judiciaire, les échantillons d'écritures manuscrites ont la même valeur que les empreintes digitales, et peuvent donc permettre l'identification des individus. Si le problème de l'identification du scripteur est souvent posé dans les courts de justice où on doit se prononcer sur l'authenticité d'un document (exemple : un testament), il peut également se poser auprès des établissements bancaires pour la vérification des signatures [Plamondon]. Enfin, dans le domaine de la génétique littéraire, des chercheurs étudient le processus de création d'un auteur à partir de l'analyse de ses manuscrits. Différents questionnements surviennent alors dont l'authentification des mains ayant pris part à la rédaction d'un document ainsi que la datation des écrits principalement.

Ces différentes problématiques ne sont pas nouvelles en elles même mais en revanche on cherche plutôt à évaluer l'apport des nouvelles technologies numériques et les avancées des méthodes de traitement automatique d'image et de reconnaissance de l'écriture à ce domaine resté confidentiel jusqu'à présent. Parallèlement, on peut rapprocher cette problématique de celle de l'identification biométrique (empreintes digitales, empreintes faciales, voix, signatures...) qui a connu ces dernières années un regain d'intérêt également. Pour toutes ces problématiques on distingue deux approches complémentaires qui sont la tâche d'identification et la tâche de vérification. Dans le domaine de l'analyse des écritures on peut les définir comme suit :

1. La tâche d'identification du scripteur consiste à identifier parmi un ensemble de scripteurs connus du système l'auteur d'un échantillon d'écriture.
2. La tâche de vérification du scripteur, quant à elle consiste à déterminer si deux échantillons d'écritures sont Oui ou Non l'œuvre de la même main.

La difficulté de la tâche d'identification du scripteur croit avec le nombre de scripteurs à différencier. Lorsque celui-ci devient important et selon la précision que l'on souhaite atteindre, la tâche d'identification du scripteur peut être considérée comme une première étape de filtrage de la base de scripteurs avant la tâche de vérification. Dans ce cas, la tâche de vérification du scripteur peut consister en une mise en correspondance entre l'échantillon du scripteur inconnu avec chaque échantillon des scripteurs du sous-ensemble sélectionné dans la base des scripteurs. De ce fait, la tâche de vérification peut parfois être adaptée pour chaque scripteur de référence connu pour une description individuelle de son écriture. Cependant, quand le nombre des scripteurs potentiel est trop grand voir inconnu ou infini, une description individuelle devient impossible et inutilisable. Dans ce cas, on préfère modéliser de façon générale la variabilité inter et intra scripteur pour pouvoir ensuite construire un système de décision [Srihari].

Dans cet article nous abordons successivement les deux tâches complémentaires que sont l'identification et la vérification du scripteur. Nous modélisons explicitement la tâche d'identification comme un processus de Recherche d'Information (RI) dans une base de documents manuscrits, nous proposons dans la première partie de ce papier d'utiliser l'un des modèles les plus connus dans le domaine de la RI : le modèle vectoriel de Salton [Salton]. Nous adaptons ce modèle initialement envisagé pour des documents électroniques textuels à des images de documents manuscrits. Pour cela, un ensemble particulier de caractéristiques est utilisé (les graphèmes) basé sur une procédure de segmentation de l'écriture manuscrite cursive et d'une étape de classification non supervisée.

Dans la seconde partie de cet article nous nous intéressons à la tâche de vérification ou d'authentification. L'approche proposée exploite les mêmes entités segmentées, les graphèmes, pour construire un test d'hypothèse basé sur un critère d'information mutuelle entre l'ensemble des caractéristiques et l'ensemble des scripteurs des deux documents considérés.

Les deux approches ont été évaluées sur deux bases de documents différentes : la première comprend 88 scripteurs et a été réalisée au sein de notre laboratoire *PSI* ; la seconde base comprend 150 scripteurs et provient de la base *IAM* réalisée à l'université de Bern en Suisse [Zimmermann]. La base *PSI* est écrite en Français tandis que la base *IAM* est rédigée en Anglais. Des performances tout à fait intéressantes par rapport à l'état de l'art ont été obtenues à l'aide de la méthodologie proposée. Nous présentons les deux approches d'identification et de vérification aux paragraphes 2 et 3 suivants.

# 2. Identification du scripteur

Si à première vue il peut sembler naturel de vouloir rapprocher la tâche d'identification du scripteur de la tâche de reconnaissance de l'écriture, une analyse plus approfondie des deux problématiques montre en fait qu'elles ne posent pas du tout les mêmes types de difficultés. En effet, si l'identification du scripteur cherche à tirer profit de la variabilité des écritures pour parvenir à son but, la tâche de reconnaissance en revanche doit parvenir à éliminer cette variabilité entre les écritures (entre les scripteurs) afin de mieux reconnaître le contenu textuel des messages. C'est la raison pour laquelle des approches sont encore à développer spécifiquement pour l'identification des scripteurs.

### 2.1. Travaux antérieurs

Les caractéristiques traditionnellement utilisées dans les approches d'identification du scripteur sont principalement des

caractéristiques globales, basées sur des mesures statistiques, extraites des blocs de texte à identifier. Ces caractéristiques peuvent être classées en deux familles :

- Caractéristiques issues de la texture : l'image du document est vue comme une simple image et non comme une écriture. Dans ce cas, l'application des filtres de Gabor et des matrices de co-occurrences ont été envisagées dans [Said].
- Caractéristiques structurelles : dans ce cas les caractéristiques extraites tentent de décrire les propriétés structurelles de l'écriture. On peut citer par exemple, la hauteur moyenne, la largeur moyenne, l'inclinaison moyenne et même la lisibilité moyenne des caractères [Marti].

Notons qu'il est également possible de combiner les deux familles de caractéristiques [Srihari]. Toutes ces approches restent toutefois difficilement comparables par manque de références communes. Cependant, nous pouvons classer les travaux précédents selon le nombre de scripteurs et la nature des échantillons d'apprentissage utilisés par le système. Un système d'identification du scripteur doit être capable de traiter le plus grand nombre de scripteurs possible, et peut selon les cas chercher à identifier un individu soit à partir de plusieurs lignes de textes soit à partir d'un seul mot manuscrit. Les travaux présentés dans [Said], par exemple, permettent d'identifier le bon scripteur, à partir de l'analyse de quelques lignes de textes dans près de 95 % des cas, sur une base comprenant 40 scripteurs. Les taux de bonne identification atteints dans les travaux de [Zois] avoisinent les 92,48 % sur une base comprenant 50 scripteurs et en utilisant 45 échantillons par scripteur d'un même mot au cours de la phase d'apprentissage. Notons que la base de scripteurs utilisée dans les travaux de [Srihari] est la plus grande qui ait été réalisée jusqu'à présent dans le domaine (1 000 scripteurs). Dans cette base chaque scripteur a été invité à recopier trois fois le même texte. Dans les travaux de [Schomaker], les auteurs ont définit un ensemble de caractéristiques sur un premier groupe composé de 100 scripteurs. La tâche d'identification a été ensuite appliquée sur un second groupe de 150 scripteurs. Un taux d'identification de 95 % a été obtenu en utilisant un même texte écrit en majuscule. Le tableau 1 ci-dessous résume les performances des différentes études.

Ce tableau met en évidence la quasi absence d'études menées dans des conditions se rapprochant de cas d'utilisation réalistes, c'est-à-dire sans contrainte de lexique (l'identification ne doit pas dépendre d'un contenu textuel spécifique) ni de style d'écriture (on ne doit pas contraindre les individus à écrire en scripte par exemple).

L'approche que nous proposons dans cet article a été conçue et évaluée sans dépendre de ces deux contraintes. Nous décrivons maintenant brièvement l'organisation du système d'identification avant de présenter en détail chacune des étapes dans les paragraphes suivants.

### 2.2. Organisation du système

La figure 1 donne un bref aperçu de la séquence de traitement des données. Elle comprend trois étapes principales : l'étape de pré-traitement, durant laquelle on s'intéresse à localiser l'information qui sera utilisée par le processus d'identification. L'étape d'extraction de caractéristiques est une phase de recherche d'une représentation pertinente des scripteurs, cette représentation est nécessaire pour la dernière phase du traitement qui est la décision. Nous allons à présent décrire chacune de ces phases (figure 1).

#### 2.2.1. Pré-traitements des documents manuscrits

Dans un premier temps, les masses connexes de l'image binarisée du document numérisé à 300 points par pouce sont extraites et analysées afin d'en éliminer certains bruits. En utilisant un seuil adaptable, déterminé à partir de la taille moyenne des masses connexes, une procédure de filtrage permet de détecter automatiquement la plupart des masses pertinentes. Les masses connexes résultantes sont ensuite segmentées en graphèmes. Les graphèmes sont des formes élémentaires de l'écriture manuscrite, produites par un algorithme de segmentation, basé sur l'analyse des minima des contours supérieurs (*cf.* figure 2.a) [Nosary]. Les minima locaux des contours supérieurs sont détectés et constituent des points de segmentation potentiels. De ces points de segmentation sont éliminés tous ceux qui coupent une occlusion. La figure 2 donne un exemple de segmentation obtenue sur le mot « manuscrit ». Les images des graphèmes segmentés sont représentées par alternance de noir et gris (*cf.* figure 2.b). La concaténation de deux (respectivement trois)

*Tableau 1. Comparaison des performances et des conditions de test d'identification du scripteur dans les travaux les plus récents*

| | # Scripteurs | Taille des échantillons et dépendance au lexique | Performances (%) |
|---|---|---|---|
| Said 2000 | 40 | Quelques lignes d'écriture manuscrite | 95 |
| Zoïs 2000 | 50 | Un seul mot en apprentissage et en test | 92,48 |
| Marti 2001 | 20 | 5 échantillons d'un même texte en apprentissage et en test | 90 |
| Srihari 2001 | 100<br>900 | Lettre du CEDAR / 1 paragraphe / 1 mot de la lettre du CEDAR | 82 / 49 / 28<br>59 / 25 / 9 |
| Schomaker 2004 | 150 | Un texte copié en lettres majuscules | 95 |

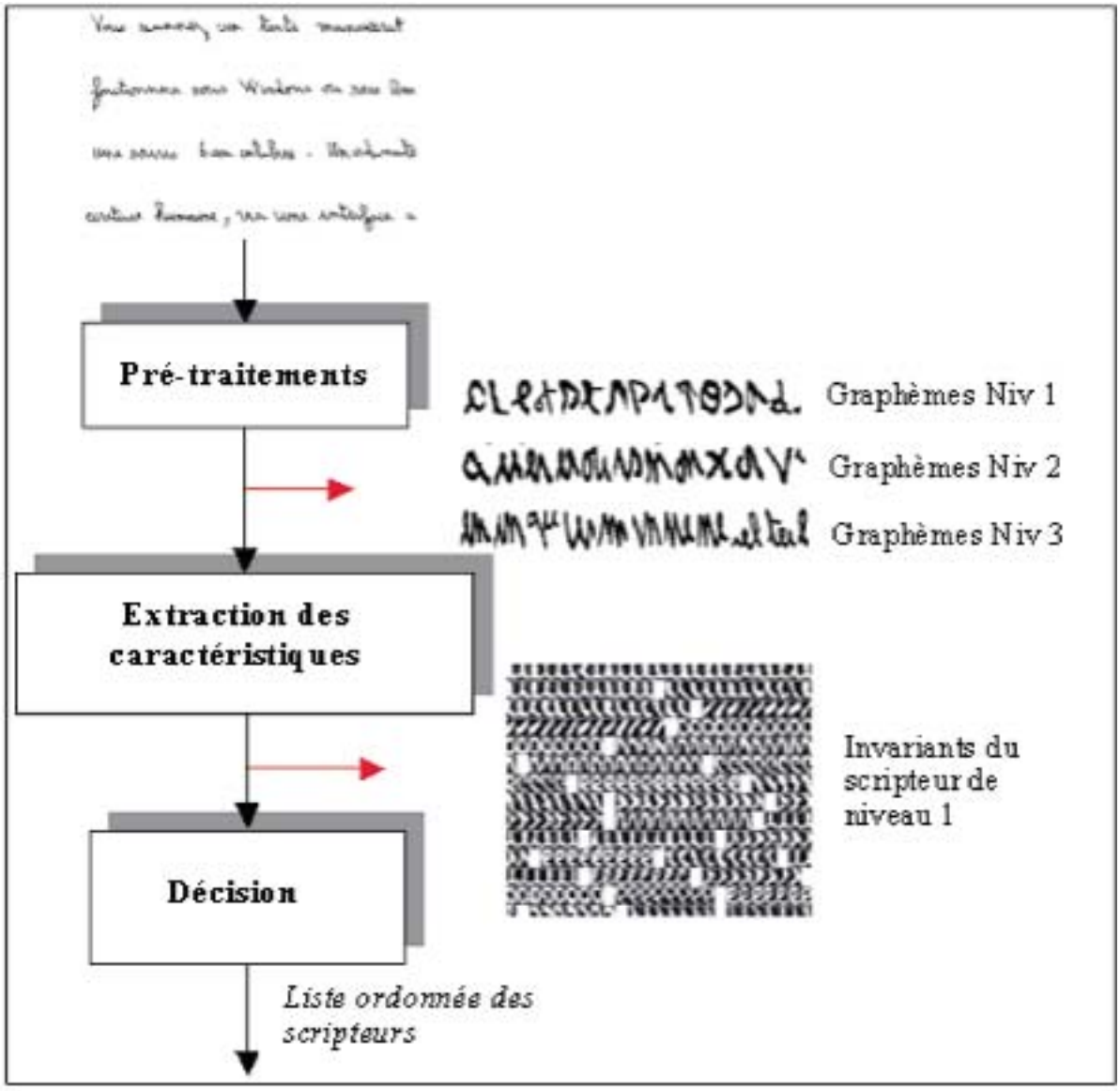



*Figure 1. Organisation du système d'identification du scripteur.*

graphèmes adjacents produits ce que nous appelons les bigrammes (respectivement trigrammes).

### 2.2.2. Extraction de caractéristiques

La tâche d'identification du scripteur réside dans la définition d'un espace de caractéristiques commun à tous les documents manuscrits. Nous recourons pour cela, à une étape de classification automatique des graphèmes segmentés qui permet ensuite de définir un ensemble de caractéristiques binaires. En effet, en affectant chaque graphème au centroïde le plus proche, on comptabilise le nombre d'occurrences de chaque centroïde dans le document analysé. De cette manière, l'ensemble de caractéristiques binaires est adapté aux écritures analysées et non pas défini *a priori*, comme c'est le cas dans l'approche proposée dans [Schomaker], où les auteurs ont défini un ensemble fixe de caractéristiques pour les scripteurs de la base d'apprentissage en utilisant une carte auto-organisatrice. Nous décrivons brièvement la procédure de classification automatique utilisée. Contrairement à la plupart des méthodes de regroupement telles que les *k*-moyennes ou les cartes auto-organisatrices, notre approche a la particularité d'adapter le nombre de clusters au problème. Elle consiste en une suite d'itérations d'une procédure de regroupement séquentiel pour converger vers les centres des classes. C'est donc une procédure simple et rapide en comparaison des approches classiques déjà citées. Elle nécessite la définition d'une mesure de similarité et d'un seuil de regroupement. Les principales étapes de cette procédure sont les suivantes :

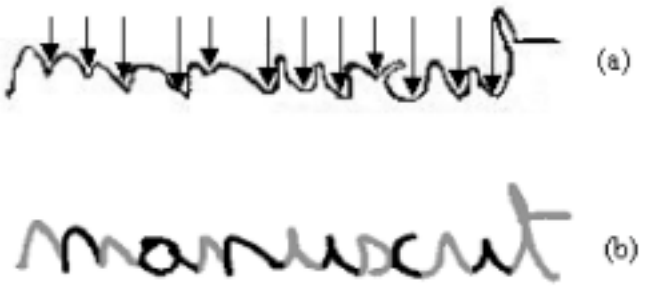



*Figure 2. Points de segmentation potentiels sur le contour supérieur de l'image (a) et les images des graphèmes résultant (b).*

```
Début
  Définir un seuil de proximité S
  Répéter le regroupement séquentiel 3 fois
    Etape 1  (initialisation)
      Choisir aléatoirement un premier candidat
      dans la base de données et le définir comme
      le premier centroïde de classe ;
    Etape 2 (Choisir le prochain candidat dans la
            base de données)
      Déterminer le centroïde le plus proche ;
      Si la distance au centroïde le plus proche
      est inférieure à S alors
        Affecter le candidat à ce centroïde;
      Sinon
        Créer un nouveau centroïde avec le candidat
        non affecté;
    Etape 3 (fin test)
      Aller à l'étape 2 jusqu'à ce que toute la
      base soit visitée
  FinRépéter
  Déterminer les intersections des clusters obtenus
  au cours des 3 itérations;
  Mettre les candidats hors intersection dans de
  nouveaux centroïdes;
Fin
```

Bien que le regroupement séquentiel soit une méthode très rapide, il est cependant très sensible à l'ordre dans lequel la base de données est parcourue. Puisque le centroïde n'est pas mis à jour lors de l'affectation de l'étape 2. C'est pourquoi nous itérons la procédure trois fois, avec un ordonnancement aléatoire des éléments de la base de données à chaque itération. Chacune des trois phases du regroupement séquentiel fourni un nombre variable de classes. Les classes invariantes (groupes invariants) sont définit comme les classes qui contiennent toujours les mêmes formes (graphèmes) durant chacune des phases du regroupement séquentiel. La figure 3 donne quelques groupes d'invariants, les plus fréquents obtenus sur la base de données *PSI*. Ces derniers constituent un ensemble de caractéristiques binaires qui sont à la base de notre méthode d'identification du

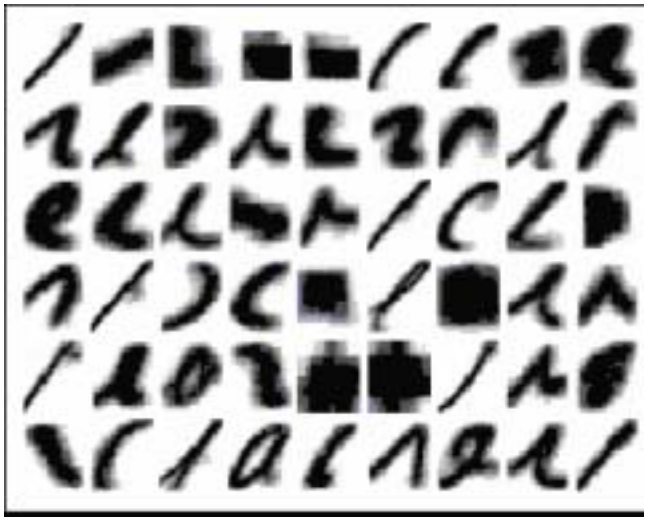


*Figure 3. Quelques invariants de niveau 1.*

*Tableau 2. Représentation du nombre des graphèmes et des classes invariantes sur les deux bases de référence.*

| | | Niveau 1 | Niveau 2 | Niveau 3 |
|---|---|---|---|---|
| Base_PSI | Nombre de graphèmes | 43178 | 25088 | 15953 |
| | Nombre de groupes d'invariants | 7230 | 13876 | 13669 |
| | Effectif du groupe le plus important | 824 | 103 | 33 |
| Base_IAM | Nombre de graphèmes | 58090 | 30456 | 17352 |
| | Nombre de groupes d'invariants | 9566 | 17878 | 14858 |
| | Effectif du groupe le plus important | 900 | 172 | 53 |

scripteur. Le nombre de clusters obtenus avec cette méthode dépend du seuil de proximité *S* ainsi que de la variabilité dans la base de données. La méthode a fourni trois mille clusters environ sur la base *IAM*. Le tableau 2 indique la taille du vecteur de caractéristiques selon la base de données analysée.

### 2.2.3. Modèle de recherche d'information

La recherche d'information est le processus de recherche, dans une base de données, des documents qui sont considérés pertinents au sens d'un besoin exprimé par l'utilisateur sous la forme d'une requête. Pour cela, la requête et les documents de la base sont généralement représentés dans un même espace de caractéristiques. De ce fait, le choix des caractéristiques est particulièrement primordial. Comme les documents doivent être décrits de façon à pouvoir répondre à tout type de requête, on ne peut en général faire intervenir une quelconque étape de sélection de caractéristiques pour réduire la dimension de l'espace et

offrir ainsi un gain en temps de calcul. Aussi cherche-t-on le plus souvent à décrire les documents en conservant l'ensemble des caractéristiques extraites et donc en ayant recours à une description dans un espace de grande dimension. Ici nous formulons explicitement le problème d'identification du scripteur comme un problème de recherche d'information afin de déterminer le ou les documents les plus pertinents au sens d'une requête graphique (ensemble de graphèmes extraits du document à identifier) dans une grande base de documents (ensemble des documents de référence). Les documents de cette base seront classés au sens d'une mesure de similarité avec la requête, du plus proche au plus éloigné.

Il existe plusieurs types de modèles de Recherche d'Information [Song] : le modèle booléen, le modèle probabiliste et le modèle vectoriel (VSM) sont les plus connus. Ce dernier, proposé par Salton [Salton], est un des modèles les plus utilisés. Bien que très simple et de conception assez ancienne, ce modèle reste très efficace [Feng]. C'est ce modèle, initialement proposé pour travailler sur des documents électroniques textuels que nous avons adapté pour travailler sur des images de textes. Nous cherchons donc à tirer pleinement profit de la particularité des textes en utilisant ce modèle.

La stratégie de recherche s'effectue en deux phases : une phase préalable d'indexation permet de décrire chaque document dans un espace de grande dimension; la phase de recherche quant à elle permet d'évaluer la pertinence de chaque document $D_j$ de la base par rapport à une requête spécifique $Q$. Cette évaluation n'est rien d'autre qu'un produit scalaire entre le vecteur décrivant la requête $Q$ et celui décrivant un document de la base $D_j$.

*1. Phase d'Indexation*

Supposons défini l'ensemble des caractéristiques. On note $\varphi_{i\ 1\leqslant i\leqslant m}$ la $i^{\text{ème}}$ caractéristique. Dans les modèles RI, chaque caractéristique peut décrire un document de la base (ou la requête) selon sa fréquence d'apparition dans ce même document, et sa fréquence d'apparition dans les autres documents de la base. Partant de ce principe chaque document de la base $D_j$ ainsi que la requête $Q$, peuvent être décrits comme suit :

$$\vec{D}_j = (a_{0j}, a_{1j}, \ldots a_{m-1j})^T \quad \text{et} \quad \vec{Q} = (b_0, b_1, \ldots b_{m-1})^T$$

où $a_{i,j}$ et $b_i$ représentent les poids attribués à chaque caractéristique $\varphi_i$, et sont définis par :

$$a_{i,j} = TF(\varphi_i, D_j) IDF(\varphi_i) \quad \text{et} \quad b_i = TF(\varphi_i, Q) IDF(\varphi_i)$$

où $TF(\varphi_i, D_j)$ est le nombre de fois où la caractéristique $\varphi_i$ apparaît dans le document $D_j$ (*Terme Frequency*). $IDF(\varphi_i)$ est l'inverse du nombre de documents possédant la caractéristique $\varphi_i$ (*Inverse Document Frequency*) ; sa valeur est donnée par :

$$IDF(\varphi_i) = \log\left(\frac{1+n}{1+DF(\varphi_i)}\right)$$

où $n$ est le nombre total de documents dans la base, et $DF(\varphi_i)$ est le nombre de documents où la caractéristique $\varphi_i$ apparaît (*Document Frequency*). Notons que si $IDF(\varphi_i) = 0$, cela signifie que la caractéristique $\varphi_i$ apparaît dans tous les documents de la base. De ce fait, cette caractéristique sera affectée d'un poids nul [Schaüble].

*2. Phase de Recherche*

On définit une mesure de similarité entre chaque document et la requête afin d'ordonner les documents selon leur pertinence. Il existe plusieurs mesures de similarité dans la littérature, la plupart d'entre elles (Dice, Jaccard, Okapi…) sont définies dans un espace de caractéristiques binaires. Lorsqu'on travaille avec des vecteurs de caractéristiques réels, on peut définir une autre mesure de similarité basée sur le produit scalaire normalisé, tel le cosinus de l'angle entre les deux vecteurs. La mesure de similarité entre un document du corpus $D_j$ et une requête $Q$ est simplement définie par :

$$\cos(Q, D_j) = \frac{\sum TFIDF_{\varphi_i, D_j} TFIDF_{\varphi_i, Q}}{\sqrt{\sum_{\varphi_i} TFIDF^2_{\varphi_i, D_j} \sum_{\varphi_i} TFIDF^2_{\varphi_i, Q}}}$$

Où les deux termes du dénominateur représentent respectivement la norme du document, et de la requête. La complexité du processus de recherche est linéaire avec la taille de la base de documents. Naturellement la taille du vecteur de caractéristiques augmente également lorsque la base de scripteurs augmente. Le tableau 2 donne un ordre de grandeur de cette augmentation : augmentation de 34 % du vecteur de caractéristiques pour près du doublement de la base de scripteurs.

### 2.3. Application

Dans cette section, nous abordons l'implémentation du modèle vectoriel pour la tâche d'identification du scripteur. Le point central réside dans la définition d'un espace commun de caractéristiques pour toute la base de données. Les phases d'indexation et de recherche peuvent ensuite être implémentées selon les étapes décrites dans la section 2.2.3. Nous avons choisi d'appliquer le principe du regroupement séquentiel sur l'ensemble des documents de toute la base de données, ce qui permet de générer un grand espace de caractéristiques (groupes d'invariants) capable de décrire tous les documents de tous les scripteurs.

Pour la phase d'évaluation deux bases de données ont été utilisées. La première a été construite au sein du laboratoire PSI, elle comprend 88 scripteurs qui ont été invités à recopier une lettre en Français comprenant 107 mots (figure 4.a). Les images des textes numérisées ont été découpées en deux parties : Deux tiers pour la base d'apprentissage et un tiers pour la base de test. La deuxième base et une fraction de la base de données IAM réalisée à l'université de Bern [Zimmermann]. Elle comprend 150

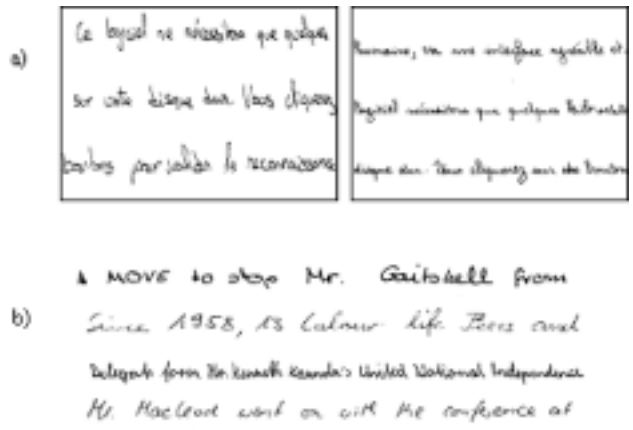


*Figure 4. Quelques échantillons de la PSI_DataBase (a) et de la IAM_DataBase (b)*

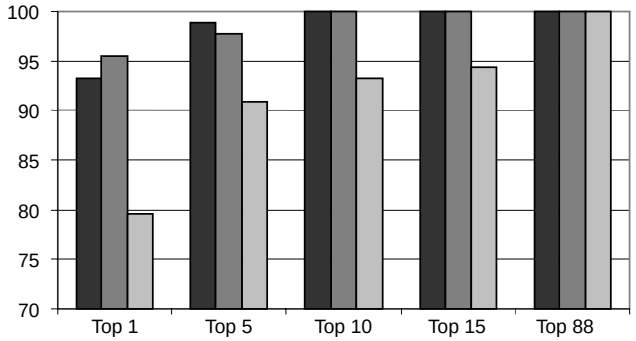
textes, écrits en Anglais, par 150 scripteurs. Le contenu textuel de ces textes diffère d'un scripteur à un autre (figure 4.b) ce qui permettra d'évaluer la tâche d'identification dans des conditions indépendantes d'un contenu textuel. Le tableau 2 donne un récapitulatif de la taille de l'espace de caractéristiques commun (vecteur de caractéristiques) pour chacune des deux bases. Sachant que les graphèmes peuvent être groupés pour former des bigrammes ou des trigrammes, l'identification du scripteur a été envisagée sur chacun des trois niveaux. En effet, il est difficile de prévoir le niveau de caractérisation que peuvent apporter ces différents niveaux d'analyse.

### 2.4. Performances

Le tableau 3 dresse un récapitulatif des résultats de notre approche sur les deux bases *PSI_DataBase* et *IAM_DataBase*. Les performances sur la base *PSI* avoisinent les 93% de bonne identification, tandis qu'avec la base IAM, ce taux est de 86%. La différence de performance entre les deux bases (10% pour les réponses de Rang 1) peut être expliquée par le faible nombre de scripteurs dans la *PSI_DataBase* (88) comparé à la *IAM_DataBase* (150). Un autre argument peut être invoqué pour expliquer les faibles performances sur la base *IAM* avec les bigrammes. En effet les échantillons utilisés dans la construction de cette dernière sont tous de taille plus petite que ceux utilisés dans la construction de la base *PSI*. Ils décrivent donc moins bien chaque écriture. En ce qui concerne les trigrammes, les performances restent relativement faibles sur les deux bases, ceci est principalement dû à la nature même des trigrammes qui sont des formes très dépendantes du contenu textuel car dans leur grande majorité elles recouvrent plus d'un caractère. On peut cependant penser que les trigrammes peuvent constituer des caractéristiques pertinentes pour certains scripteurs, mais ces formes restent peut fréquentes dans un texte de taille raisonnable (un paragraphe de quelques lignes) et sont par conséquent beaucoup moins exploitables. Comme première conclusion, ces résultats permettent de dire que le modèle vectoriel est pertinent dans l'identification du scripteur, en utilisant des caractéristiques locales.

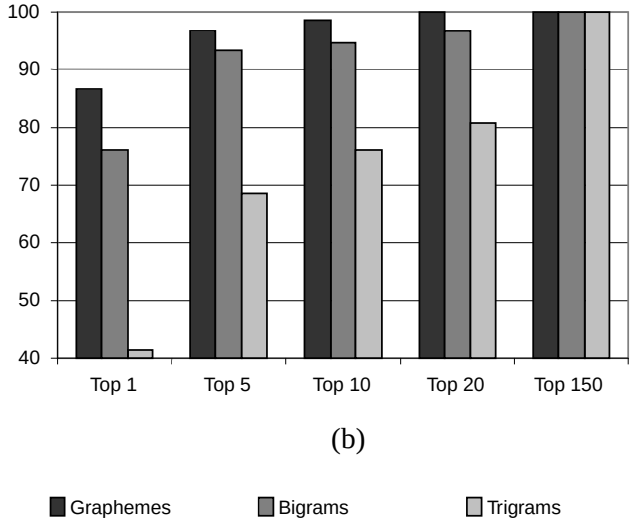
*Tableau 3. Performances de l'identification du scripteur sur la PSI_DataBase (a) et sur la IAM_DataBase (b)*

| ***PSI*** | Top 1 | Top 5 | Top 10 | Top 15 |
|---|---|---|---|---|
| Graphèmes | 93,18 | 98,86 | 100 | 100 |
| Bigrams | 95,45 | 97,72 | 100 | 100 |
| Trigrams | 79,54 | 90,9 | 93,18 | 94,31 |




(a)

| ***IAM*** | Top 1 | Top 5 | Top 10 | Top 20 |
|---|---|---|---|---|
| Graphèmes | 86,66 | 96,66 | 98,66 | 100 |
| Bigrams | 76 | 93,33 | 94,66 | 96,66 |
| Trigrams | 41,33 | 68,66 | 76 | 80,66 |

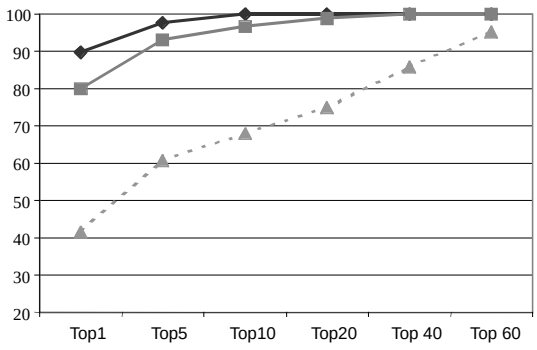



(b)

Une deuxième série de tests a été conduite sur les mêmes bases, afin d'évaluer l'influence de la taille de la requête sur les performances d'identification (voir tableau 4). En utilisant une séquence de 50 graphèmes, un taux de 90% de bonne identification a été obtenue sur la base *PSI*. Ces performances chutent à 68% sur la base *IAM*, dans les mêmes conditions. Dans les deux cas, on a pu remarquer que les trigrammes ont un pouvoir

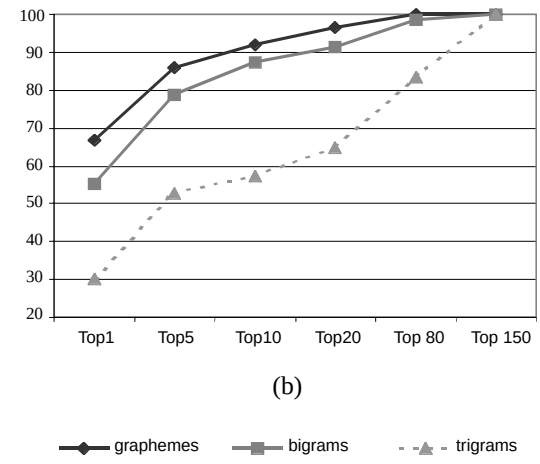
*Tableau 4. Performances d'identification du scripteur avec des requêtes courtes de 50 graphèmes Sur la base PSI (a) st sur la base IAM (b).*

| ***PSI*** | Top1 | Top5 | Top10 | Top20 | Top40 |
|---|---|---|---|---|---|
| Graphèmes | 89,77 | 97,72 | 100 | 100 | 100 |
| Bigrams | 79,88 | 93,14 | 96,72 | 98,92 | 100 |
| Trigrams | 41,66 | 60,7 | 67,85 | 75 | 85,7 |




(a)

| ***IAM*** | Top1 | Top5 | Top10 | Top20 | Top 80 |
|---|---|---|---|---|---|
| Graphèmes | 66,67 | 86 | 92 | 96,66 | 100 |
| Bigrams | 55,33 | 78,66 | 87,33 | 91,33 | 98,66 |
| Trigrams | 30 | 52,66 | 57,33 | 64,66 | 83,33 |




(b)

discriminant moins significatif que celui des graphèmes ou des bigrammes, comme nous l'avons déjà remarqué dans la première série de tests tableau 3.

### 2.5. Discussion

Les premiers résultats montrent que le Modèle Vectoriel est pertinent pour la tâche d'identification du scripteur, en utilisant des caractéristiques locales. En outre, les bigrammes peuvent être de meilleures caractéristiques pour cette tâche. Cette approche qui ne dépend ni du contenu textuel ni ne contraint les scripteurs à un style d'écriture particulier a été testée sur deux bases différentes de taille significative en comparaison aux travaux les plus récents dans le domaine. Bien que testées dans des conditions plus générales, les performances obtenues par notre approche sont du même ordre voire meilleures que les travaux de la littérature.

Toutefois, certaines limitations générales à toutes les approches d'identification peuvent être mises en avant. En effet, l'identification du scripteur étant basée sur le principe de similarité entre documents et requête, la sortie de ce processus est une liste ordonnée des documents de la base de référence. Ce principe atteint sa limite si le scripteur à identifier est inconnu. Dans ce cas, il faudrait être en mesure de rejeter la requête. L'approche de vérification du scripteur (ou authentification) apporte une solution à cette problématique dans la mesure où elle permet de rendre une décision du type acceptation/rejet. Nous proposons dans la partie suivante une approche de vérification du scripteur basée également sur l'utilisation des graphèmes.

# 3. Vérification du Scripteur

La tâche de vérification du scripteur consiste à analyser deux documents manuscrits, dont le(s) scripteur(s) ne sont pas forcément connus par le système, et de décider si oui ou non ils ont été écrits par la même main. La plupart du temps cette tâche est réalisée par un expert et est empreinte d'une subjectivité importante. En tout état de cause, la confiance que l'on peut associer à une décision de ce type n'est pas scientifiquement démontrée. Des travaux récents ont proposé une méthodologie d'analyse scientifique des écritures pour la tâche de vérification du scripteur [Cha]. Il faut noter que cette tâche a été rarement étudiée au regard des nombreuses études qui ont concerné la tâche d'identification. Ceci est sans doute dû au fait que la vérification implique des processus de décisions locales qui dépendent généralement du contenu textuel. En effet, on se ramène généralement à devoir comparer les formes possibles d'un caractère ou d'un mot spécifique présent sur les documents étudiés. C'est pourquoi l'automatisation complète de cette tâche semble peu réaliste car dépendant en premier lieu de la tâche de reconnaissance automatique.

Nous proposons dans cette communication d'aborder la tâche de vérification du scripteur en restant indépendant du contenu textuel. Ceci n'est possible que lorsque l'information disponible pour l'analyse est suffisamment conséquente c'est-à-dire dans les mêmes conditions que pour la tâche d'identification du scripteur (bloc de texte). Si cette orientation peut sembler, *a priori*, très limitative pour l'expertise, elle apparaît néanmoins

tout à fait complémentaire de l'étape d'identification du scripteur que nous venons de présenter dans la section 2. En effet, comme nous l'avons déjà souligné, par construction l'approche d'identification ne permet pas de détecter une écriture inconnue de la base de données. L'approche de vérification qui est proposée va permettre de valider ou de rejeter les documents manuscrits retournés par la phase d'identification. L'approche reprend la démarche déjà exploitée lors de la phase d'identification fondée sur les graphèmes pour décider si deux documents manuscrits proviennent oui/non de la même main. Dans cette optique, nous construisons un test d'hypothèse fondé sur l'information mutuelle entre deux documents manuscrits.

### 3.1. Construction d'un test d'hypotheses

#### 3.1.1. Critère d'Information Mutuelle

Supposons que 2 documents manuscrits $D_1$ et $D_2$ aient été écrits respectivement par les scripteurs $S_1$ et $S_2$. Notons $S$ l'ensemble de ces 2 scripteurs :

$$S = \{S_1, S_2\}.$$

Nous supposons également qu'une étape de pré-traitements a permis de segmenter les deux documents manuscrits en graphèmes et qu'une étape de classification automatique (section 2.2.2) a permis de définir un ensemble $G$ de caractéristiques communes aux 2 documents analysés (des graphèmes):

$$G = \{g_1, g_2, g_3, \ldots g_N\}.$$

Notons que contrairement à la tâche d'identification, la procédure de regroupement est appliquée sur les deux documents en présence qu'ils soient ou non connus du système. Notre approche de vérification travaille donc sur un ensemble de caractéristiques adapté à chaque paire de documents analysée. Certaines de ces caractéristiques peuvent être présentes sur les deux documents, tandis que d'autres peuvent apparaître spécifiquement sur un seul document. L'information mutuelle permet alors de mesurer l'indépendance entre l'ensemble des 2 scripteurs $S$ et l'ensemble de caractéristiques $G$. De faibles valeurs de l'information mutuelle indiquent une indépendance importante entre les 2 variables aléatoires $G$ et $S$ tandis que des valeurs importantes traduisent une dépendance forte entre $S$ et $G$. L'indépendance entre $S$ et $G$ devrait indiquer que l'ensemble de caractéristiques $G$ est distribué de la même façon sur les deux documents et devrait refléter la même identité pour les 2 scripteurs $S_1$ et $S_2$. Dans le cas contraire, le critère d'information mutuelle devrait permettre de détecter une dépendance forte entre $S$ et $G$ et révéler en cela une identité différente des deux scripteurs. Nous rappelons l'expression de l'information mutuelle entre $G$ et $S$ :

$$I_M(G,S) = H(G) - H(G/S) \tag{1}$$

où $H(G)$ désigne l'entropie de Shannon [Shannon 84]:

$$H(G) = -\sum_{i=1}^{\mathrm{card}(G)} P(g_i)\log_2(P(g_i)) \tag{2}$$

et $H(G/S)$ est l'entropie conditionnelle définie par:

$$H(G|S) = -\sum_{i=1}^{\mathrm{card}(G)}\sum_{j=1}^{\mathrm{card}(S)} P(S_j)P(g_i|S_j)\log_2[P(g_i|S_j)] \tag{3}$$

soit finalement :

$$\begin{aligned} H(G|S) =& H(G|S_1) + H(G|S_2) \\ =& -\sum_{i=1}^{\mathrm{card}(G)} P(S_1)P(g|S_1)\log_2[P(g|S_1)] \\ & -\sum_{i=1}^{\mathrm{card}(G)} P(S_2)P(g|S_2)\log_2[P(g|S_2)] \end{aligned} \tag{4}$$

Les trois quantités $H(G)$, $H(G/S_1)$ et $H(G/S_2)$ sont tout d'abord calculées avant d'en déduir la valeur de l'Information Mutuelle.

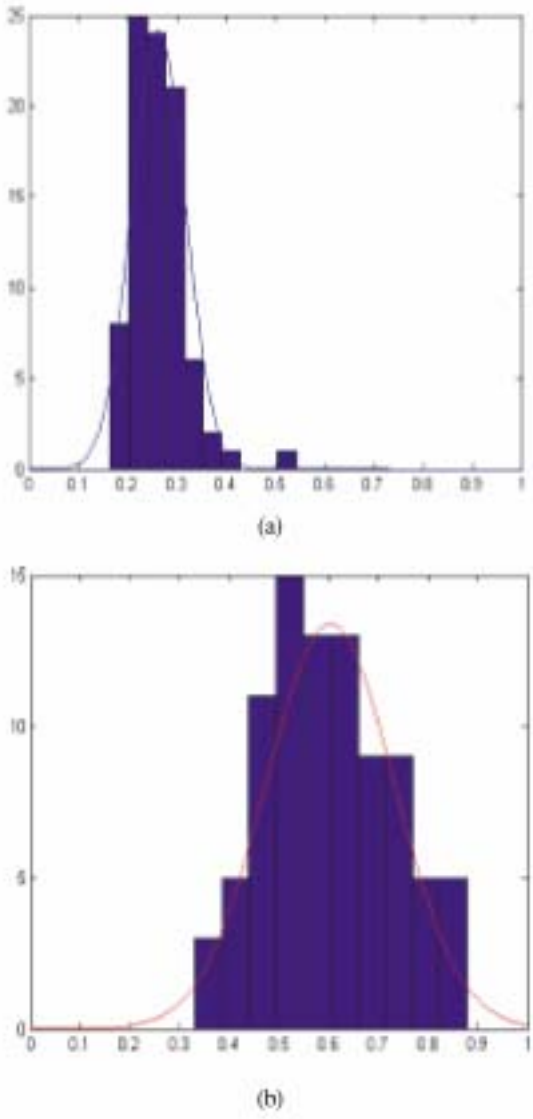



*Figure 5. Distribution du critère d'information mutuelle sous l'hypothèse $H_0$ (a) et $H_1$ (b)*

Pour attester de la pertinence de ce critère nous avons mené un premier test sur la base du *PSI*. Rappelons que chacun des textes a été divisé en deux de façon à disposer de 2 échantillons d'écriture pour chaque scripteur. La figure 5 représente la distribution intra-scripteurs du critère d'information mutuelle c'est-à-dire lorsque les deux scripteurs sont identiques (a) et la distribution inter-scripteurs (b). Ces distributions montrent clairement l'aptitude du critère d'information mutuelle à constituer un critère quantitatif pour détecter les deux situations et valider ou rejeter l'hypothèse d'identité des scripteurs des deux documents. De plus, cette figure montre que les deux distributions peuvent être ajustées approximativement par une loi normale.

#### 3.1.2. Test d'Hypothèses

Nous cherchons maintenant à construire un critère de décision entre les 2 hypothèses suivantes :

$H_0 : S_1 = S_2 \quad$ et $\quad H_1 : S_1 \neq S_2$

Pour cela, il est nécessaire de construire un test d'hypothèse [Saporta]. Si $H_0$ représente l'hypothèse nulle ou l'hypothèse par défaut. Chacune des 2 décisions est affectée d'une probabilité de bonne et de mauvaise décision (tableau 5). Traditionnellement, $\alpha$ désigne l'erreur de première espèce. C'est la probabilité d'accepter $H_1$ quand $H_0$ est vraie. Tandis que $\beta$ désigne l'erreur de seconde espèce.

*Tableau 5. Probabilités associées aux différentes décisions.*

| Vérité / Décision | $H_0$ est vraie | $H_1$ est vraie |
|---|---|---|
| Accepter $H_0$ | $1-\alpha$ | $\beta$ |
| Accepter $H_1$ | $\alpha$ | $1-\beta$ |

En faisant l'hypothèse de normalité des deux distributions du critère d'information mutuelle, il est très simple de quantifier les erreurs de première et de seconde espèce. Le seuil optimal de décision $V_c$ a été choisi au point d'intersection des deux distributions. La zone de rejet de $H_0$, notée $W_0$, est définie par l'erreur de première espèce. La limite de cette zone, correspondant à la valeur $V_c$, permet de déterminer la zone de rejet de l'hypothèse $H_1$, notée $W_1$ et de déduire ainsi l'erreur de seconde espèce par :

$P(W_0|H_0) = \alpha \quad$ et $\quad P(W_1|H_1) = \beta$

De la même manière, on peut déterminer les régions d'acceptations des deux hypothèses, pour $H_0$ et pour $H_1$. On a :

$P(\overline{W}_0|H_0) = 1-\alpha \quad$ et $\quad P(\overline{W}_1|H_1) = 1-\beta$

### 3.2. Expérimentation

Nous avons évalué cette approche de vérification des scripteurs sur la base *IAM*, mais c'est la base *PSI* qui a servi à déterminer les régions d'acceptation des deux hypothèses en déterminant la valeur optimale du seuil $V_c$. Le tableau 6 résume les résultats obtenus sur la *PSI_DataBase*.
Le test a été ensuite appliqué sur des couples de scripteurs choisis au hasard dans la base *IAM*. Rappelons que cette base d'écritures a été constituée par des scripteurs Suisses et rédigée en langue anglaise [Zimmermann]. Cette seconde base est donc *a priori* très différente de la base *PSI*. Le test de vérification des scripteurs sur cette seconde base a permis d'obtenir les résultats présentés dans le tableau 7.

*Tableau 6. Erreur de première espèce, puissance du test et valeur du seuil sur la PSI_DataBase.*

| *PSI_DataBase* | $\alpha$ | $1-\beta$ | $V_c$ |
|---|---|---|---|
| Gramphèmes | 3,5 % | 97,5 % | 0,36164 |
| Bigrams | 3,4 % | 98,5 % | 0,4984 |

*Tableau 7. Performances de la vérification du scripteur sur la base IAM pour les valeurs du seuil $V_c$ données au tableau 6.*

| *IAM_DataBase* | $\alpha$ | $1-\beta$ |
|---|---|---|
| Gramphèmes | 4 % | 96 % |
| Bigrams | 10,66 % | 96,66 % |

### 3.3. Discussion

Pour ce qui concerne spécifiquement l'approche proposée dans cette section pour la vérification des scripteurs, les résultats semblent particulièrement prometteurs à plusieurs égards. Tout d'abord le choix d'une représentation locale fondée sur les graphèmes segmentés semble tout à fait pertinent puisqu'il permet un niveau de description proche des caractères sans toutefois nécessiter une étape de reconnaissance. Nous parvenons donc à vérifier des écritures en étant indépendant d'un contenu textuel. D'autre part, il est remarquable d'obtenir un niveau de performances sur la base *IAM* du même ordre que sur la base *PSI* sur laquelle a été effectué l'apprentissage du test d'hypothèses. Nous sommes donc en mesure d'apporter des éléments quantitatifs pertinents quant à l'hypothèse d'individualité de l'écriture. Nous montrons de plus ici qu'il est possible de construire un test statistique robuste sur plusieurs bases d'écritures.

# 4. Conclusion

Dans cet article nous avons présenté 2 approches complémentaires pour la reconnaissance automatique du scripteur. D'une part nous avons adapté et appliqué une approche de recherche d'information, traditionnellement réservée aux documents électroniques. La technique développée apporte une réponse au problème de l'identification du scripteur d'un document et offre un potentiel important d'extension sur de grandes bases de documents patrimoniaux par exemple. Outre son utilisation spécifique sur des écritures manuscrites, cette technique pourrait facilement être étendue à d'autres problèmes de caractérisation de documents textuels par leurs contenus graphiques. Citons par exemple les problèmes d'identification de typographies sur des documents imprimés anciens. Notons également que l'approche est par construction compatible avec les techniques de compression à base de dictionnaires de formes telles que celle utilisée par les normes JBIG ou DjVu. Pour toutes ces raisons, la technique semble particulièrement intéressante.

D'autre part nous avons proposé un test d'hypothèse permettant de vérifier la compatibilité entre les écritures de 2 documents différents. Cette étape de vérification du scripteur est indispensable pour valider les hypothèses émises par le système d'identification proposé précédemment. L'approche montre des capacités de vérification excellentes tant en acceptation qu'en rejet et une aptitude à être généralisée sur des écritures inconnues très prometteuse. L'approche envisagée ici dans le cadre complémentaire d'une technique de recherche d'information pourrait tout à fait se concevoir dans le contexte de l'identification biométrique ou de l'expertise judiciaire. Dans cette perspective il conviendrait toutefois d'évaluer l'approche sur des exemples de contrefaçon. Néanmoins la méthodologie proposée nécessite de travailler sur des échantillons de taille suffisante pour être indépendante des contenus textuels. Une approche spécifique reste à développer pour travailler sur des échantillons de faible taille.

# Références


[Cha] S.H. CHA, S. SRIHARI, « Multiple Feature Integration for Writer Verification », 7th International Workshop on Frontiers in Handwriting Recognition, Amsterdam, pp. 333-342, 2000.

[Feng] D. FENG, W.C. SIU, H.J. ZHANG, *Multimedia Information Retrieval and Management*, Springer Edition, 2003.

[Marti] U.V. MARTI, R. MESSERLI, H. BUNKE, « Writer Identification Using Text Line Based Features », Proc. ICDAR'01, Seattle (USA), pp. 101-105, 2001.

[Nosary] A. NOSARY, *Automatic recognition of handwritten texts through writer adaptation*, PhD dissertation (in french), University of Rouen, France, 2002.

[Plamondon] R. PLAMONDON, G. LORETTE, « Automatic signature verification and writer identification – the state of the art », Pattern Recognition, vol. 22, n°2, pp. 107-131, 1989.

[Said] H.E.S. SAID, T.N TAN, K.D. BAKER, « Personal Identification Based on Handwritting », Pattern Recognition, vol. 33, pp. 149-160, 2000.

[Salton] G. SALTON, WRONG, « A vector Space Model for Automatic Indexing », *Information Retrieval and Language Processing*, pp. 613-620, 1975.

[Saporta] G. SAPORTA, *Probabilités analyse des données et statistiques*, *Edition Technip*, pp. 317-330, 1990.

[Shannon] C. SHANNON, The Mathematical Theory of Communication. Bell System Technical. Journal, Roberts, J.A. vol 27, pp. 379-423, 1984.

[Schaüble] P. SCHAÜBLE, *Multimedia Information Retrieval : Content-Based Information Retrieval From Large Text and Audio Databases*, Kluwer Academic Publishers, 1997.

[Schomaker] L. SCHOMAKER, M. BULACU, Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercases Western Script, IEEE-PAMI, vol. 26, N° 6, pp. 787-798, 2004.

[Song] F. SONG, W. BRUCE CROFT, « A General Language Model for Information Retrieval », Eighth International Conference on Information and Knowledge Management (ICIKM'99), 1999.

[Srihari] S. SRIHARI, S.H. CHA, H. ARORA, S. LEE, « Individuality of Handwritig : A Validity Study », Proc. ICDAR'01, pp.106-109, 2001.

[Zimmermann] M. ZIMMERMANN, H. BUNKE, « Automatic Segmentation of the IAM Off-line Handwritten {English} Text Database », 16th International Conference on Pattern Recognition, Vol 4, pp. 35-39, 2002.

[Zois] E.N. ZOIS, V. ANASTASSOPOULOS, « Morphological Waveform Coding for Writer Identification », Pattern Recognition, vol. 33, n°3, pp.385-398, 2000.