Identification et Vérification du Scripteur : 2 approches complémentaires pour l'analyse quantitative des documents manuscrits

Writer Identification and Verification : 2 complementary approaches for the quantitative analysis of handwritten documents

A. Bensefia, T. Paquet et L. Heutte

Laboratoire PSI - FRE CNRS 2645
UFR des Sciences, Université de Rouen
76821 Mont-Saint-Aignan Cedex, France
Thierry.Paquet@univ-rouen.fr

Résumé

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.

Mots Clef

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

Abstract

In this paper, we apply an Information Retrieval model for the writer identification task. A set of local features is defined by clustering the graphemes produced by a segmentation procedure. Then a textual based Information Retrieval model is applied. After a first indexation step, this model no longer requires image access to the database for responding to a specific query, thus making the process particularly effective. Image queries are handwritten documents projected in the feature space prior to the retrieval of the suitable responses. Writer hypothesis retrieved are then analysed during a verification phase. We have called upon a mutual information criterion to verify that two documents may have been produced by the same writer or not. Hypothesis testing is used for this purpose. The method is tested on two different databases and appears to be robust over these two databases thus making the approach very promising for large scale applications in the domain of handwritten document querying.

Keywords

Handwritten documents, writer identification, writer verification, information retrieval, mutual information, hypothesis testing.

1 Introduction

La préservation du patrimoine culturel a depuis fort longtemps imposé une certaine limitation des accès aux biens. L'évolution des technologies numériques permet d'envisager aujourd'hui des campagnes de numérisation des collections à grande échelle. Pour le grand public, il deviendra dès lors possible dans un avenir proche d'accéder aux fonds numérisés dans un but de consultation d'archives ou de parcours culturel. Pour les institutions cette perspective laisse entrevoir la possibilité de valoriser les collections auprès du public en les rendant enfin accessibles. Les verrous scientifiques qui limitent encore actuellement la faisabilité de tels projets sont liés à la quantité d'images qu'il faut pouvoir analyser, décrire, indexer automatiquement et de manière robuste pour permettre des accès pertinents aux contenus. Pour un public de spécialistes, experts littéraires, conservateurs…, la numérisation facilitera également l'accès aux archives mais elle leur permettra surtout de bénéficier de l'apport des techniques d'analyse numérique, de traitement d'images et de reconnaissance de formes pour développer de nouveaux moyens d'analyse scientifique des collections numérisées.

Dans cet article, nous nous intéressons plus particulièrement à la problématique de l'analyse des documents manuscrits pour l'expertise. La démarche que

nous proposons permet d'envisager un accès aux documents manuscrits numérisés selon leur modalité graphique c'est-à-dire en fonction des écritures et des auteurs sur lesquels l'utilisateur porte son intérêt. En effet, pour ces documents, on peut globalement distinguer deux types d'utilisation auxquelles on peut associer des requêtes de nature très différente :

- Les archives de documents manuscrits peuvent être vues sous l'angle de leurs contenus textuels. Dans ce cas l'interrogation des bases documentaires nécessite de recourir à une phase préalable de transcription des textes manuscrits permettant ensuite une analyse textuelle. L'état de l'art dans ce domaine ne permet pas d'envisager des applications immédiates, la reconnaissance de l'écriture restant en effet mal maîtrisée sur des applications omni-scripteurs et faisant appel à des lexiques de grande taille [6].
- Les archives de documents manuscrits peuvent également être vues sous l'angle de leurs contenus graphiques. Dans ce cas l'interrogation des bases documentaires est effectuée à partir de requêtes graphiques. On cherche par exemple à retrouver les documents de la base présentant certaines calligraphies correspondant à certains scripteurs. D'autres cas d'utilisation peuvent concerner la détection des différentes mains présentes, ou bien la datation des documents par rapport à la chronologie de l'œuvre de l'auteur. Les techniques de traitement de l'écriture permettent d'envisager dès à présent certaines applications comme celle que nous proposons dans cette communication.

On peut considérer que ces deux cas d'utilisation relèvent d'un problème de recherche d'information soit textuelle soit graphique. Ces deux tâches ont été largement étudiées soit dans le domaine documentaire soit en traitement d'images. En ce qui concerne spécifiquement l'analyse des écritures manuscrites, cette tâche relève de l'identification du scripteur d'un document. Un certain nombre de travaux ont abordé ce problème, le plus souvent en s'appuyant sur des techniques d'analyse de textures pour caractériser les écritures. L'originalité des travaux que nous présentons tient au fait que nous fondons notre démarche sur une technique de recherche d'information textuelle mais en utilisant une description spécifique à l'écriture manuscrite et à l'image.

Dans la première partie de cette communication, nous présentons l'approche retenue pour la tâche d'identification du scripteur: les caractéristiques utilisées sont les entités graphiques segmentées par un algorithme de découpage en graphèmes des composantes connexes des textes manuscrits ; l'approche s'inspire du modèle vectoriel de recherche d'information de Salton [9] initialement développé pour la recherche d'information textuelle. Nous justifions le bien fondé de ce choix en évaluant la tâche d'identification du scripteur sur une base d'écritures constituée au laboratoire comportant 88 scripteurs. Nous nous intéressons également à évaluer l'influence de la taille de la requête sur la qualité des réponses fournies par le système.

Dans la seconde partie de cette communication, nous développons une analyse complémentaire qui a pour but de valider les résultats de la première approche. Elle concerne la tâche de vérification ou encore d'authentification du scripteur. En effet si l'approche proposée pour l'identification du scripteur est pertinente pour retrouver des écritures connues du système, elle est en revanche incapable de détecter des requêtes constituées d'écritures inconnues et dans ce cas de rejeter toute proposition de scripteur. L'approche proposée exploite les entités segmentées, les graphèmes, pour construire un test d'hypothèse fondée sur l'information mutuelle entre les textes manuscrits. L'approche est évaluée sur deux bases d'écritures différentes : la base construite au laboratoire composée de 88 scripteurs et une seconde comportant près de 150 scripteurs et qui a été rédigée en langue anglaise [14]. Les résultats obtenus montrent d'excellentes capacités à authentifier les écritures d'un même scripteur pour des requêtes de taille importante.

2 Identification du Scripteur

Pour l'approche d'identification, l'utilisateur dispose d'un seul document qui constitue la requête, et cherche à identifier son auteur parmi un ensemble de N scripteurs connus. Bien que l'identification du scripteur s'inscrive dans la même problématique que la reconnaissance de l'écriture, elle ne semble pas, cependant, poser le même type de difficultés. En effet, la tâche d'identification peut tirer profit de la variabilité des écritures afin de les discriminer, tandis que la tâche de reconnaissance doit au contraire parvenir à s'affranchir de la variabilité entre les scripteurs pour identifier le message textuel quel qu'en soit le scripteur.

2.1 Travaux antérieurs

Jusqu'à présent, les caractéristiques utilisées dans ces deux approches sont des caractéristiques globales qui utilisent des mesures statistiques, extraites de l'ensemble du bloc de texte à identifier. Ces caractéristiques se classent en deux types:

- *caractéristiques issues de la texture :* l'image du document est vue dans ce cas simplement comme une image et non comme une écriture. Par exemple, l'application des filtres de Gabor et des matrices de cooccurrences a été envisagée dans [8].
- *caractéristiques structurelles :* dans ce cas les caractéristiques extraites s'attachent à décrire les particularités de l'écriture. On peut citer pour l'exemple des caractéristiques telles que la hauteur moyenne, la largeur moyenne, l'inclinaison moyenne et la lisibilité moyenne des caractères [4].

Il est également possible de combiner les deux groupes de caractéristiques [7].

Ces caractéristiques de nature statistique extraites d'un bloc de texte permettent d'atteindre des résultats intéressants, qu'il est toutefois toujours délicat de comparer par manque de références communes.
Finalement, on peut catégoriser les différents travaux proposés d'une part selon le nombre de scripteurs à discriminer et d'autre part selon la taille de l'échantillon disponible pour procéder à l'identification du scripteur (un ensemble de lignes de texte ou au contraire quelques mots). Ainsi par exemple les travaux proposés dans [8] permettent d'identifier 95% des 40 scripteurs que le système peut traiter à partir d'un échantillon de quelques lignes d'écriture. Les travaux présentés dans [9] permettent d'identifier le bon scripteur dans 92,48% des cas parmi 50 scripteurs en utilisant 45 échantillons du même mot que les participants ont été invités à écrire. Il faut noter que le travail proposé dans [7] a traité le problème de l'identification/vérification du scripteur avec le plus grand des corpus (1000 scripteurs), en utilisant une base de données constituée d'un même texte écrit 3 fois par chaque scripteur.

2.2 Organisation du système

Nous présentons dans les paragraphes qui suivent les différentes étapes nécessaires à l'identification du scripteur. La figure 1 donne un bref aperçu de la chaîne de traitement. Elle fait intervenir trois étapes classiques en reconnaissance des formes : une étape de pré-traitements dont l'objectif principal est de localiser les informations, une étape d'extraction de caractéristiques dont le but est d'obtenir une représentation pertinente pour la prise de décision qui représente l'étape finale de la chaîne de traitement.

2.3 Pré-traitements des documents

Les constituants de l'écriture (masses connexes) sont tout d'abord analysés afin d'éliminer certaines représentations graphiques comme les ratures, les zones de surcharge ou sous-lignées dont on sait *a priori* qu'elles ne caractérisent pas l'écriture. Chaque entité retenue est ensuite segmentée en lettres ou en morceaux de lettres que nous appelons : *graphèmes*. Cette dénomination ne se réfère à aucune description spécifique de l'écriture et il est admis qu'elle peut porter à confusion. Les graphèmes sont en effet des formes élémentaires de l'écriture manuscrite au sens d'un algorithme de segmentation [6]. La concaténation de deux graphèmes adjacents (respectivement 3) donne ce que nous appelons les graphèmes du second niveau ou bi-grammes (respectivement graphèmes du troisième niveau ou tri-grammes).

2.4 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 recourant à une description dans un espace de grande dimension.
On peut formuler notre problème d'identification du scripteur comme un procédé de recherche par le contenu 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 [13] : le modèle booléen, le modèle probabiliste et le modèle vectoriel (VSM) sont les plus connus. Ce dernier, proposé par Salton [9], est un des modèles les plus utilisés. Bien que très simple et de conception assez ancienne, ce modèle reste très efficace [5] [7].
Dans ce modèle, la stratégie de recherche s'effectue en deux phases : une phase préalable d'indexation permet de décrire chaque document par un vecteur de grande dimension; la phase de recherche 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.

Phase d'indexation

Supposons défini l'ensemble des caractéristiques. On note $\varphi_i \ 1 \le i \le m$ la $i^{ème}$ 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 :

$$\vec{D}_j = (a_{o,j},\ a_{1,j},\ldots.\qquad a_{m-1,j})^T$$

$$\vec{Q} = (b_o,\ b_1,\ \ldots.\qquad b_{m-1})^T$$

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

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

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

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

où n est le nombre total de documents dans la base, et $DF(\varphi_i)$ est le nombre de documents où la caractéristique φ_i apparaît (*Document Frequency*). Notons que si $IDF(\varphi_i) = 0$, cela signifie que la caractéristique φ_i apparaît dans tous les documents de la base. De ce fait, cette caractéristique sera affectée d'un poids nul, et devrait même être retirée de l'ensemble des caractéristiques [12].

Phase de Recherche

Chaque document ainsi que la requête est décrit dans le même espace de caractéristiques : une mesure de similarité entre chaque document et la requête est nécessaire afin d'ordonner les documents selon leur pertinence. Plusieurs mesures de similarité ont été proposées dans la littérature [2]: Dice, Jaccard, Okapi. D'après [5] la plupart des mesures de similarité ne sont que des variantes de la mesure *Cosinus*, qui consiste à calculer la valeur de l'angle entre le vecteur d'un document du corpus D_j et le vecteur de la requête Q. Elle est 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. Les scores des TF-IDF sont calculés pour chaque document durant la phase d'indexation.

2.5 Application

Le point central de l'implémentation d'un modèle de recherche d'information pour une identification du scripteur 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 précédemment. Une étape de classification automatique permet de définir un espace de caractéristiques discrètes commun à l'ensemble de la base de documents manuscrits.

La figure 2 présente quelques groupes d'invariants obtenus sur la base de référence, ces caractéristiques peuvent apparaître chez plusieurs scripteurs. Au sens du modèle, une caractéristique est considérée comme non pertinente si elle est partagée par tous les scripteurs.

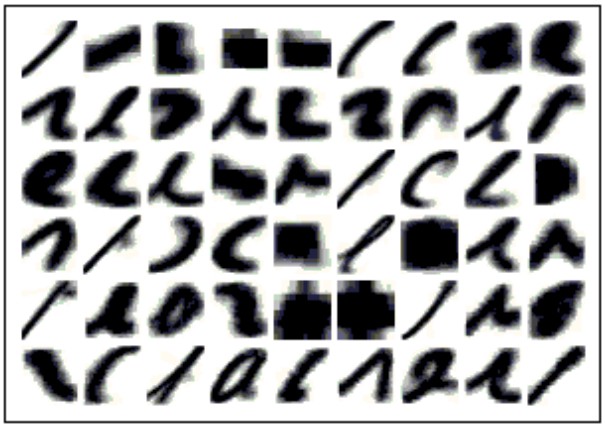


Figure 2. Quelques groupes d'invariants de niveau 1, de la base du laboratoire

Evaluation sur des requêtes longues

Le tableau 1 donne le résultat de l'approche sur notre base. On constate que le bon scripteur a été correctement identifié dans près de 93% des cas (83/88), en ayant recours aux graphèmes du premier niveau. Ce taux d'identification monte jusqu'à 95.45% (84/88) avec les bi-grammes (graphèmes niveau 2), quand les tri-grammes ne donnent que 80% (70/88) de bonne identification.

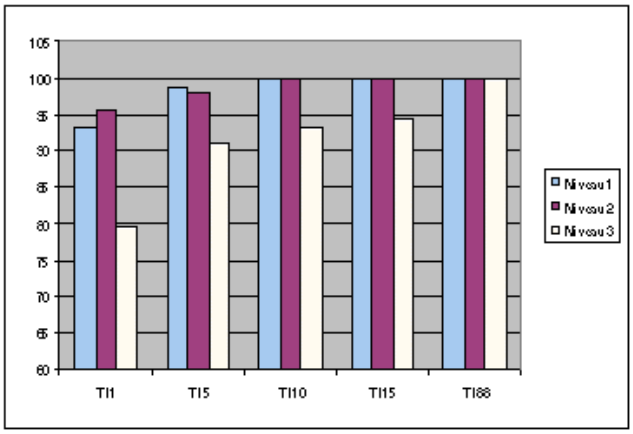


Tableau 1 : Identification du scripteur sur des requêtes longues.

Evaluation sur des requêtes courtes

Notre seconde série de tests a porté sur l'évaluation de l'approche du modèle vectoriel avec des requêtes courtes d'une longueur de: 10, 20, 30, 40 et 50 graphèmes du niveau 1. Des résultats présentés dans le tableau 2, apparaît l'efficacité de l'approche avec les requêtes courtes, car avec des séquences de 50 graphèmes (équivalent à 3 ou 4 mots), on est arrivé à près de 93% de taux de bonne identification en 1er choix (Top1) et 100% dans les 5 premiers choix (Top5).

Ces résultats, ainsi que ceux présentés dans le tableau 1, mettent en évidence l'influence de la taille de l'échantillon à identifier, dans un procédé d'identification du scripteur.

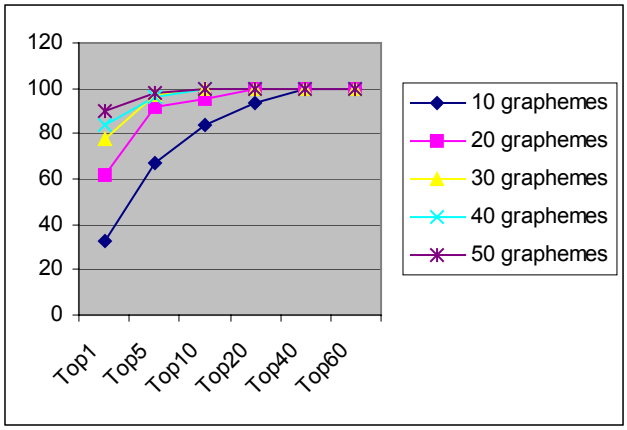


Tableau 2 : Identification du scripteur sur des requêtes courtes.

2.6 Discussion

L'identification du scripteur est basée sur le principe de similitude entre la requête (document à identifier) et tous les documents de la base de référence. La sortie de ce processus est une liste ordonnée des documents de la base. Ce principe soulève le problème où le scripteur à identifier est inconnu de la base de référence. Dans ce cas, une solution possible réside dans l'authentification (vérification) du scripteur proposé par l'approche d'identification. En effet avec une approche de vérification, nous pouvons accepter ou rejeter le scripteur proposé avec un certain taux d'erreur.

3. Vérification du scripteur

La tâche de vérification du scripteur consiste à authentifier le scripteur d'un document. 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 notamment pour la tâche de vérification du scripteur [3]. 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. 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. En effet, par construction, l'approche ne permet pas de détecter une écriture inconnue de la base de donnée. L'approche de vérification qui est proposée permet de valider ou de rejeter les documents manuscrits retournés par la phase d'identification.

L'approche met à profit l'ensemble des caractéristiques locales déjà exploitées lors de la phase d'identification (les graphèmes), pour attester que deux documents manuscrits peuvent provenir de la même main ou non. A cette fin, nous construisons un test d'hypothèse fondé sur l'information mutuelle entre deux documents manuscrits.

3.1 Construction d'un test d'hypothèse

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 des 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. De plus, une étape de classification automatique (comme dans le cas de l'identification) permet de définir un ensemble G de caractéristiques communes aux 2 documents analysés : $G = \{ g_1, g_2, g_3, \ldots. g_N \}$. 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. Des faibles valeurs de l'information mutuelle indiquent une indépendance importante entre les 2 variables aléatoires tandis que les 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 l'identité différente des deux scripteurs.

Nous rappelons l'expression de l'information mutuelle entre G et S:

$$I_M(G,S) = H(G) - H(G/S) \qquad (1)$$

Où $H(G)$ désigne l'entropie de Shannon [11] :

$$H(G) = -\sum_{i=1}^{card(G)} P(g_i)\, H(G=g_i) \qquad (2)$$

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

$$H(G|S) = -\sum P(Sj) H(G|S=Sj) - \sum_{i=1}^{card(G)} \sum_{j=1}^{card(S)} P(S_j)\, P(g|S_j) \log_2 \left[P(g|S_j) \right] \qquad (3)$$

Pour attester la pertinence de ce critère nous avons mené un premier test sur la base des 88 scripteurs (base du laboratoire). Rappelons que chaque texte a été partagé en deux de façon à disposer de 2 échantillons d'écriture pour chaque scripteur. La figure 3 représente la distribution du critère d'information mutuelle en intra-scripteurs c'est-à-dire lorsque les deux scripteurs sont identiques (a) et en inter-scripteurs c'est-à-dire lorsque les deux scripteurs sont différents (b). Ces distributions montrent clairement l'aptitude du critère d'information mutuelle à constituer un critère quantitatif pour valider ou rejeter l'hypothèse d'identité des scripteurs des deux documents.

Test d'hypothèse

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

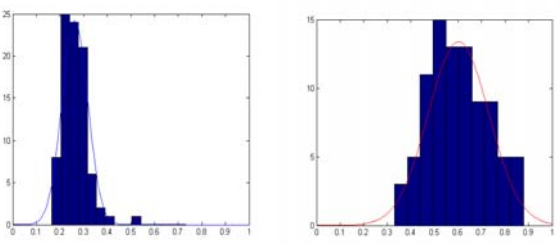
$$H_0 : S_1 = S_2 \quad \text{et} \quad H_1 : S_1 \neq S_2 \qquad (4)$$

Pour cela il s'agit de construire un test d'hypothèse [10]. H_0 représente l'hypothèse nulle ou hypothèse par défaut. Chacune des 2 décisions est affectée d'une probabilité de bonne et mauvaise décision (tableau 3). Classiquement, α désigne l'erreur de première espèce. C'est la probabilité d'accepter H_1 quand H_0 est vraie. Tandis que β désigne l'erreur de seconde espèce. C'est la probabilité d'accepter H_0 quand H_1 est vraie.

Vérité / Décision	H_0 est vraie	H_1 est vraie
accepter H_0	$1-\alpha$	β
accepter H_1	α	$1-\beta$

Tableau 3: Probabilités associées aux différentes décisions.

En faisant l'hypothèse de normalité de la distribution du critère d'information mutuelle dans les deux situations où les hypothèses sont vérifiées, il est très simple de quantifier les erreurs de première et seconde espèce.
A l'aide de ces distributions et en s'imposant une valeur pour l'erreur de première espèce on peut définir les régions d'acceptation et de rejet des deux hypothèses et déduire la valeur expérimentale de β .



(a) **(b)**

Figure 3. Distribution du critère d'information sous l'hypothèse H_0 (a) et H_1 (b)

La région de rejet de H_0 , notée W_0, est définie par l'erreur de première espèce. La limite de cette région permet de définir la région de rejet de H_1, notée W_1 et de déduire l'erreur de seconde espèce par les relations suivantes :

$$P(W_0 \mid H_0) = \alpha \quad \text{et} \quad P(W_1 \mid H_1) = \beta \qquad (5)$$

De la même manière (figure 4), on détermine les régions d'acceptation des 2 hypothèses soit $\overline{W_0}$ pour la région d'acceptation de H_0 et $\overline{W_1}$ pour la région d'acceptation de H_1. Nous avons :

$$P(\overline{W_0} \mid H_0) = 1-\alpha \quad \text{et} \quad P(\overline{W_1} \mid H_1) = 1 - \beta \qquad (6)$$

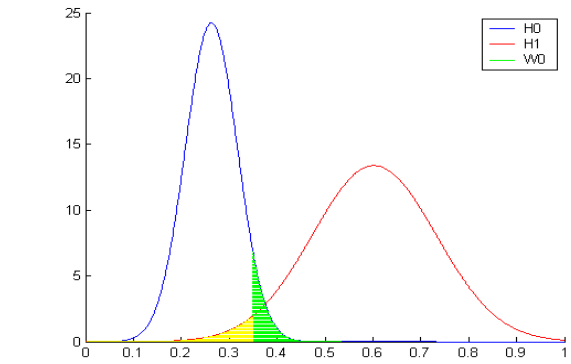


Figure 4. Régions de rejet de chacune des 2 hypothèses.

3.2 Expérimentation

Nous avons évalué ce test de vérification des scripteurs sur les deux bases d'écritures déjà utilisées. Dans un premier temps la base du laboratoire a permis de fixer les régions d'acceptation des 2 hypothèses en imposant une valeur à l'erreur de première espèce. Ceci permet dans un premier temps d'évaluer la puissance du test (1-β) ainsi construit à accepter l'hypothèse H_1 (tableau 4).

α	1 - β
0,05	0.9744
0,025	0.9641

Tableau 4: Erreur de première espèce et puissance du test.

Dans un second temps le test a été appliqué sur des couples de scripteurs choisis au hasard sur la seconde base d'écriture. Rappelons que cette seconde base d'écritures a été constituée par des scripteurs suisses et rédigée en langue anglaise [14]. Cette seconde base est

donc *a priori* très différente de la première. Le test de vérification des scripteurs sur cette seconde base a permis d'obtenir les résultats présentés dans le tableau 5.

Bonne acceptation		Bon Rejet	
$\alpha = 5\%$	$\alpha = 2.5\%$	$\alpha = 5\%$	$\alpha = 2.5\%$
94%	97,33%	97.33%	94%

Tableau 5 : Performances en vérification sur la base Suisse.

Nous avons dans un second temps évalué l'aptitude du test à vérifier 2 scripteurs à partir d'échantillons de petite taille comme nous avons pu le faire pour la tâche d'identification. Le tableau 6 donne l'aptitude du test proposé à prendre la bonne décision parmi les 2 hypothèses possibles.

Taille de la Requête	Bonne acceptation		Bon Rejet	
	$\alpha = 5\%$	$\alpha = 2.5\%$	$\alpha = 5\%$	$\alpha = 2.5\%$
50	82.66%	88%	75.33%	68%
40	90%	94%	56%	47,33%
30	92.66%	94.66%	31.33%	24.66%
20	98.66%	99.33%	9.33%	7.33%
10	100%	100%	0.6%	0%

Tableau 6 : Performances en vérification sur des requêtes courtes.

3.3 Discussion

Pour ce qui concerne spécifiquement l'approche proposée dans ce paragraphe pour la vérification des scripteurs, les résultats semblent particulièrement prometteurs à plusieurs égards.

Tout d'abord le choix d'une représentation locale fondées 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.

D'autre part, il est remarquable d'obtenir un niveau de performance sur la base Suisse du même ordre que sur la base PSI sur laquelle a été effectué l'apprentissage du test d'hypothèse. 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 construir un test statistique robuste sur plusieur bases d'écritures. Il conviendra naturellement de valider l'approche sur des bases de documents plus conséquentes.

Toutefois les tests effectués sur des requêtes courtes montrent les limites de l'approche. En effet sur des échantillons de petite taille l'approche devient trop dépendante du contenu textuel. Contenu dont nous n'avons absolument pas tenu compte. Il conviendrait dans ce cas de développer spécifiquement une approche qui ne soit pas biaisée par les contenus textuels différents des deux documents analysés.

4. Conclusion

Dans cet article nous avons présenté 2 approches complémentaires pour l'analyse quantitative des documents manuscrits.

D'une part nous avons proposé une approche de recherche d'information compatible avec les spécificités des documents manuscrits intégrant les 2 modalités sous-jacentes à ces documents : modalités textuelle et graphique. 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.

Remerciements

Ce travail s'inscrit dans le cadre du programme STIC-SHS Société de l'information du CNRS en collaboration avec l'ITEM (Institut des Textes et Manuscrits Modernes).

Bibliographie

[1] Bensefia A., Nosary A., Paquet T., Heutte L., Writer Identification by Writer's Invariants, International Workshop on Frontiers in Handwriting Recognition, Niagara on the Lake, Canada, pp 274-279, 2002.
[2] Duda R., Stork D., Hart P., Pattern Classification and Scene Analysis, Wiley & Sons, 2nd Edition, 2000.
[3] Cha, S.H., & Srihari, S. (2000). Multiple Feature Integration for Writer Verification. *Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition; IWFHR VII* . (pp.333-342).
[4] Marti U.V., Messerli R., Bunke H., *Writer Identification Using Text Line Based Features*, International Conference on Document Analysis and Recognition, Seattle, USA, pp. 101-105, 2001.
[5] Memmi D., *Le modèle vectoriel pour le traitement de documents*, Les Cahiers du Laboratoire Leibniz, IMAG-Grenoble, France, n°14, 2000.
[6] Nosary A., Reconnaissance Automatique de Textes Manuscrits par Adaptation au Scripteur, Thèse de Doctorat, Université de Rouen, 2002.
[7] Pouliquen B., Delamane D., Lebeux P., « Indexation des textes médicaux par extraction de concepts et ses utilisations », *6éme Journée internationale d'Analyse statistique des Donnés Textuelles* (JADT'02), 2002.
[8] Said H.E.S., Tan T.N., Baker K.D., "Personal Identification Based on Handwritting", *Pattern Recognition*, vol. 33, pp. 149-160, 2000.
[9] Salton, Wrong, "A vector Space Model for Automatic Indexing", *Information Retrieval and Language Processing*, pp. 613-620, 1975.
[10] Saporta., G. (1990). Probabilités analyse des données et statistiques. *Edition Technip*. pp 317-330.
[11] Shannon, C. (1984). The Mathematical Theory of Communication. *Bell System Technical Journal*, Roberts, J.A. *27, 379-423.*
[12] Schaüble P., *Multimedia Information Retrieval : Content–Based Information Retrieval from Large Text and Audio Databases*, Institut de Technologie (ETH), Zurich, Suisse, Kluwer Academic Publishers, 1997.
[13] Song F., Bruce Croft W., "A General Language Model for Information Retrieval", *Eighth International Conference on Information and Knowledge Management* (ICIKM'99), 1999.
[14] Zimmermann M., Bunke, H. " Automatic Segmentation of the IAM Off-line Handwritten {English} Text Database". 16th International Conference on Pattern Recognition, Canada, Vol 4, pp. 35-39, 2002.

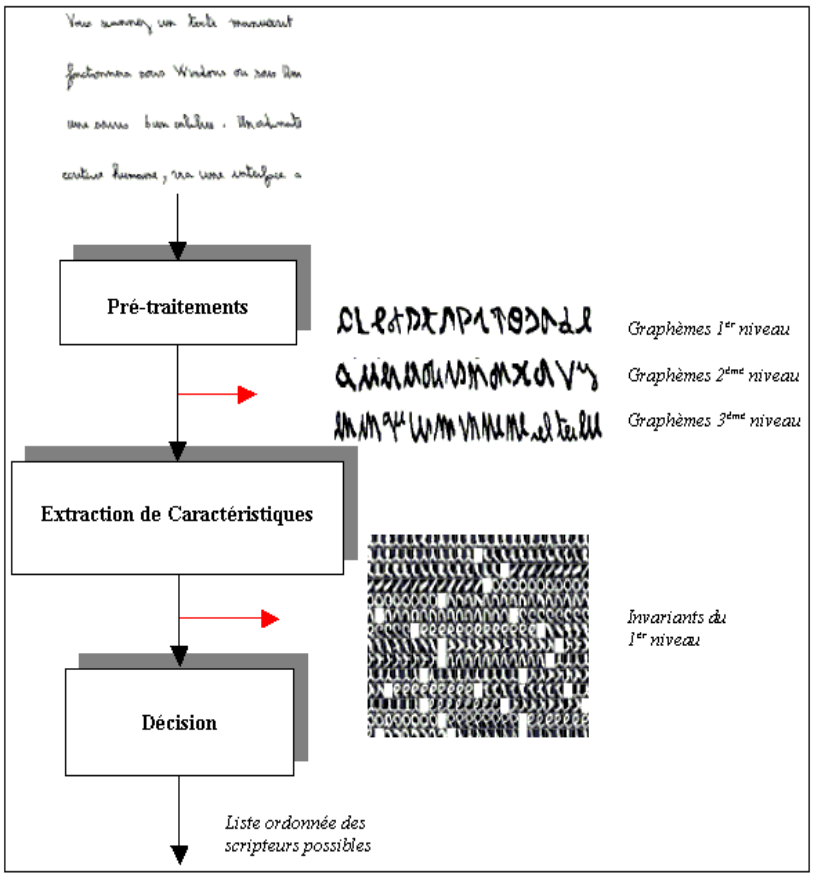


Figure 1. Organisation du système d'identification du scripteur.