

## Discrimination des styles d'écriture des manuscrits médiévaux pour la Paléographie

Ikram Moalla, Frank Lebourgeois, Hubert Emptoz, Adel Alimi

► **To cite this version:**

Ikram Moalla, Frank Lebourgeois, Hubert Emptoz, Adel Alimi. Discrimination des styles d'écriture des manuscrits médiévaux pour la Paléographie. Laurence Likforman-Sulem. Colloque International Francophone sur l'Écrit et le Document, Sep 2006, Fribourg, Suisse. SDN06, pp.223-228, 2006. <hal-00119841>

**HAL Id: hal-00119841**

**<https://hal.archives-ouvertes.fr/hal-00119841>**

Submitted on 12 Dec 2006

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Discrimination des styles d'écriture des manuscrits médiévaux pour la Paléographie

Ikram Moalla<sup>1,2</sup>, Franck LeBourgeois<sup>1</sup>, Hubert Emptoz<sup>1</sup>, Adel M. Alimi<sup>2</sup>

<sup>1</sup>Laboratoire d'InfoRmatique en Image et Systèmes d'information  
I.N.S.A. de LYON

20 Avenue A. Einstein 69621 Villeurbanne, Cedex FRANCE

<sup>2</sup>REsearch Group on Intelligent Machines  
ENIS, Sfax – Tunisie, Route de Sokra km 3

Email: ikram.moalla@ieee.org

**Résumé :** *Ce papier constitue une première contribution à la discrimination des écritures des manuscrits médiévaux pour la Paléographie. Nous cherchons à retrouver les classifications construites par les paléographes sur la généalogie des écritures et de leurs évolutions durant le moyen âge. Ce travail devrait contribuer à confirmer objectivement les travaux des paléographes et tester les possibilités de l'analyse des images dans la discrimination des écritures médiévales. Nous avons choisi de caractériser statistiquement les formes des écritures sans segmenter l'image. Nous utiliserons principalement la notion de cooccurrence comme mesure d'auto-similarité. Les premiers résultats obtenus semblent confirmer les classifications données par les experts.*

**Mots-clés :** documents anciens, paléographie, reconnaissance de styles, autosimilarité, matrice de cooccurrence.

## 1 Introduction

Dans le domaine de la recherche sur les textes anciens, la philologie (science qui s'intéresse au problème de datation, de localisation et d'édition de texte), étudie autant la manière dont les textes sont écrits ou imprimés que leurs contenus.

L'analyse des styles d'écritures apporte des informations complémentaires aux contenus des textes que l'on considère comme étant des méta-données. La manière dont le texte est représenté constitue une information introduite de façon consciente ou inconsciente par l'auteur ou le scribe qui peut permettre par exemple, de dater, d'authentifier ou d'indexer un document.

La présentation d'un document imprimé se manifeste par sa structure physique et la typographie des caractères (polices, taille, déclinaison, fonte) alors que la présentation d'un manuscrit ancien recèle d'autres

niveaux d'interprétation comme le style personnel d'écriture du scribe, la calligraphie utilisée et la mise en page du document. Ces derniers peuvent être représentatifs d'une époque et d'un lieu et servir à la datation et la transcription.

L'objet de ce travail consiste à apporter une première contribution méthodologique et applicative à l'analyse automatique des mises en forme des documents, au service de la recherche en histoire des textes. Nous nous intéresserons plus précisément aux manuscrits anciens latins du Moyen Age, période qui précède la Renaissance et l'avènement de l'imprimerie. La définition du style est multiple et complexe. Nous nous concentrerons sur une approche visuelle et perceptive du style des écritures, celle que l'on pourra qualifier et étudier avec des outils d'analyse d'images. La principale difficulté consiste à discerner le style d'une écriture manuscrite qui soit relié à la période historique et/ou une localisation géographique indépendamment du style personnel du scribe.

## 2 Définitions et généralités

La paléographie est l'étude de l'écriture et de la forme des caractères. Ses objectifs sont principalement l'enseignement du déchiffrement correct des écritures anciennes et l'étude de l'histoire de l'évolution de l'écriture. Depuis la fin du I<sup>er</sup> siècle avant J.-C., les écritures se sont transformées selon les usages mais le fonctionnement est resté le même. Du VIII<sup>e</sup> au XII<sup>e</sup> siècle, la caroline règne sur l'Occident. Elle évolue vers des formes anguleuses pour donner naissance à l'écriture gothique. Le passage d'une écriture à l'autre ne s'est pas toujours effectué de façon radicale mais par évolution lente et progressive. Ce qui explique qu'il est difficile d'identifier catégoriquement une écriture donnée. Par exemple, on observe des textes en écriture *caroline* qui contiennent déjà des attributs de l'écriture *gothique*. Le paléographe doit alors quantifier

précisément la part de mélange des familles d'écritures. C'est le cas, par exemple, de la classe d'écriture *prégothique* qui est intermédiaire entre l'écriture caroline et l'écriture gothique (FIG.1) ou encore l'écriture *Hybrida* entre la *cursiva* et la *textualis* (FIG.2). La diversification des familles d'écritures en Europe s'accélère jusqu'à la Renaissance et voit se développer des sous-familles d'écritures à l'intérieur de chaque grande famille de gothiques.

Ainsi, nous pouvons distinguer plusieurs sous-familles de *gothique cursiva* représentées dans la FIG.3 qui traduisent la précision de l'exécution de cette écriture. De même, la FIG.4 montre plusieurs sous-familles de *gothique textualis*.

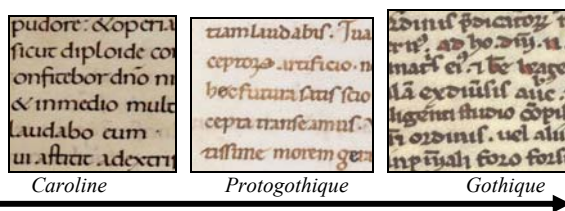


FIG.1 - Evolution progressive de la caroline à la prégothique puis à la gothique.



FIG.2 - Evolution progressive de la Gothique cursiva à la gothique hybrida puis à la gothique textualis.

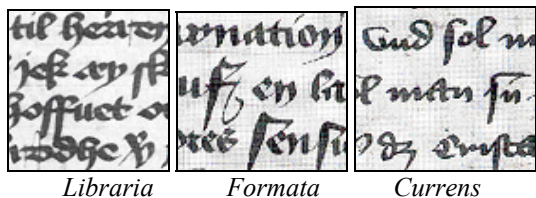


FIG.3 - Exemples de sous-familles de styles cursiva entre le VIII<sup>ème</sup> et le XVI<sup>ème</sup> siècle [DER 03]

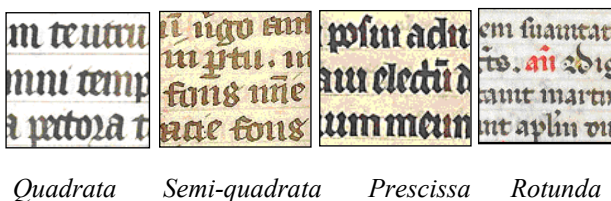


FIG.4 - Exemples de sous-familles de styles textualis entre le VIII<sup>ème</sup> et le XVI<sup>ème</sup> siècle

La FIG.5 montre la variabilité des écritures à l'intérieur d'une même sous-famille comme la classe *gothique textualis rotunda*.

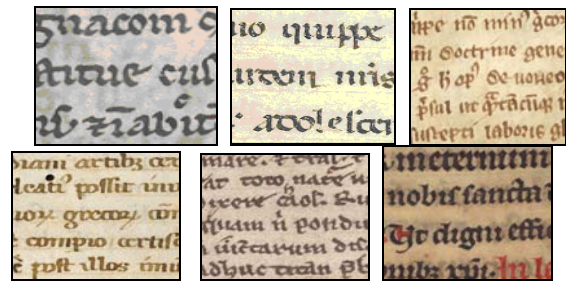


FIG.5 - Exemples d'images de textes représentant la variation intra-classe du style gothique textualis rotunda

Elle illustre la difficulté en terme d'analyse d'images à définir des descripteurs de formes pour trouver une homogénéité entre les différents échantillons d'une même écriture. La paléographie est, de ce fait, une science complexe. De plus, il existe une diversité de classifications en familles et sous-familles d'écritures qui dépendent des paléographes. Un style d'écriture peut être étiqueté différemment. Ceci évoque le caractère subjectif de la paléographie qui n'a pas de règles précises pour trouver le style exact d'une écriture. De plus, en examinant un style, le paléographe peut créer une confusion entre ses connaissances *a priori* et ses connaissances paléographiques sur les styles. Trouver un système pour assister les paléographes est donc loin d'être un travail simple.

### 3 Etat de l'art

Les travaux sur la caractérisation des écritures ont été réalisés pour des applications différentes de celle de la paléographie. Parmi ces applications nous pouvons citer la vérification et l'authentification de scripteurs, la pré-classification des écritures en terme de lisibilité pour une meilleure reconnaissance dans le tri automatique du courrier et des chèques. Toutes ces études sont connexes à notre problématique mais ces contributions ne sont pas toutes directement ré-exploitable pour l'étude paléographique. La rose des directions binaire a été utilisée par [CRE 95] pour identifier les formes différentes d'écritures en vue de leur reconnaissance. L'analyse fractale mesure les autosimilarités présentes dans une image, c'est une bonne mesure du style d'un scripteur qui peut servir à classer les écritures suivant leur lisibilité [VIN 00]. L'indice fractal est aussi susceptible de caractériser les différents alphabets dans les textes imprimés. [EGL 98] caractérise différents styles de texte par des mesures de complexité des formes, de lisibilité et de compacité indépendamment de l'alphabet utilisé. Enfin nous signalons d'autres travaux susceptibles d'être réutilisés pour la reconnaissance des écritures médiévales comme la reconnaissance des scripts (des mots dans un alphabet particulier) dans les documents multilingues. Ces travaux utilisent la similarité de graphèmes [MOA 02], la texture [TAN 98], ou l'analyse de profil de projection [WOO 95] etc.

Du point de vue analytique, nous distinguons deux approches complémentaires pour le traitement des styles d'écritures:

- **Approche locale** : elle consiste à reproduire le travail des paléographes, en cherchant à établir des similarités

visuelles entre des écritures à partir de lettres très particulières caractéristiques d'une écriture (exemples : 'r', 's', 'e', 'a') [AIO 99], [CIU 05], [YOS04]. D'autres travaux ont utilisé des données paléographiques pour des systèmes d'authentications de manuscrits sans pour autant faire une étude paléographique [LIK 90].

- **Approche globale** : elle consiste à analyser statistiquement l'image entière d'un manuscrit, sans segmenter l'image ni en lignes de texte, les mots ou les caractères. Elle consiste également à trouver des descripteurs de formes capables de distinguer les différentes écritures. Avec l'analyse globale on ne cherche pas à reproduire le travail des paléographes, mais à utiliser une méthode plus appropriée à l'analyse automatique d'images.

## 4 Méthode proposée

Notre objectif est le développement de méthodologies et d'outils d'analyse d'images pour assister les historiens dans la classification et la datation des manuscrits anciens latins à partir de la reconnaissance des écritures. En effet, chaque époque de l'histoire a été marquée par un ou plusieurs types d'écritures. Ainsi, la connaissance de l'écriture d'un document permet de connaître sa date et/ou son origine géographique. Notre domaine d'étude couvre les écritures anciennes latines du VIII<sup>ème</sup> siècle jusqu'à la Renaissance au XVI<sup>ème</sup> siècle. L'étude des écritures latines antérieures au VIII<sup>ème</sup> siècle comme l'*onciale* ou l'écriture *cursiva* n'a pas un réel intérêt pour les paléographes. En revanche l'assistance à l'expertise des écritures médiévales est très utile à partir du XII<sup>ème</sup> siècle. Il s'agit de différencier les grandes familles d'écritures comme le montre la répartition de la FIG.6.

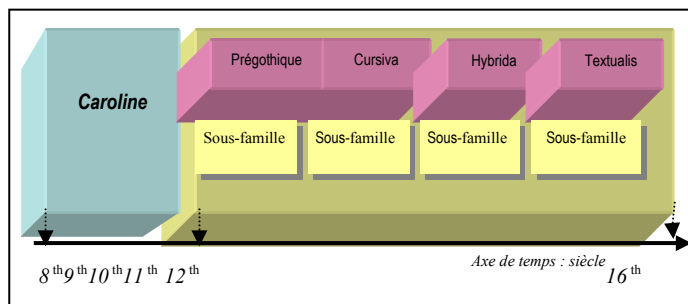


FIG. 6 - Répartition des différentes familles et sous-familles de styles latin entre le 8<sup>ème</sup> et le 16<sup>ème</sup> siècle

Dans un premier temps, notre travail s'est focalisé sur l'extraction de caractéristiques suffisamment discriminantes pour pouvoir différencier le plus grand nombre d'écritures latines possibles. Cette étude a permis d'étudier la faisabilité d'un système d'analyse automatique des images à l'usage des paléographes.

Dans un deuxième temps nous avons affiné l'étude en tentant progressivement de discriminer les deux grandes familles de styles d'écritures latines : *caroline*, *gothique* puis les sous familles de *gothiques* : *cursiva*, *hybrida* et *textualis* et enfin les sous-familles telles que la *rotunda*, la *quadrata*, la *semi-quadrata* et la *prescissa* pour la *textualis*; la *formata*, la *libraria* et la *currens* pour la

*gothique cursiva*. Cette analyse a pour objectif à la fois d'augmenter la précision de la discrimination entre les écritures et d'étudier plus en détail les confusions possibles entre les écritures proches.

### 4.1 Des conditions difficiles

Le développement d'un système d'assistance à l'expertise des manuscrits anciens est une tâche rendue difficile par de nombreux facteurs la complexité des formes d'écritures (FIG.2, 3, 4), la grande variabilité des écritures d'une même classe (FIG.5), l'existence d'écritures hybrides issues de mélanges de plusieurs écritures (FIG.1, 2), la faible qualité de conservation des manuscrits, le vieillissement des supports et des encres (FIG.7a), l'enchevêtrement des lignes et des mots (FIG.7b), la présence de notes dans la marge et/ou entre les lignes (FIG.7c) et la grande variabilité de la qualité des images de différentes origines. Certaines images en couleurs proviennent d'une numérisation de qualité, d'autres images de mauvaise qualité sont issues de la numérisation de livres ou de microfilms, en niveaux de gris. La plupart des images présentent des dégradations dues à une trop forte compression (JPEG). Il est à signaler que nos échantillons ont été numérisés avec des résolutions toutes différentes.

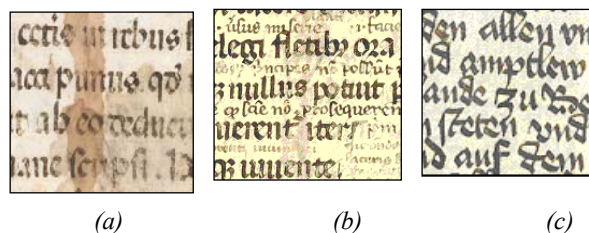


FIG.7 - mauvaise qualité des images, (a) Vieillesse de l'encre, (b) Enchevêtrement des lignes, (c) Écriture dans la marge et/ou entre les lignes [BGB]

Notre base paléographique est issue de différentes provenances [Drolez03], [BV], [BGB].

Dans ce contexte difficile, nous allons analyser directement l'image en niveaux de gris sans filtrage préalable, sans restauration et sans correction géométrique. Ce choix nous prive d'une grande partie des travaux réutilisables et en particulier de tous ceux basés sur la segmentation.

### 4.2 Notre approche

Nous avons choisi l'*approche globale* étant données les conditions difficiles décrites précédemment. Nous avons cherché des descripteurs de formes capables de distinguer les différentes écritures. Ces mesures globales devraient être indépendantes du *contenu du texte*, du *style personnel du scripteur*, de la *langue utilisée*, des *lettres employées* et de leurs fréquences. L'analyse globale s'affranchit des formes fantaisistes des caractères de début et de fin de mots, ainsi que de l'éventuelle présence d'ornements incrustés dans le texte. Ces avantages sont très précieux pour pouvoir analyser une grande quantité d'images de manuscrits de qualité très variable ainsi que les manuscrits dégradés.

Sans travaux antérieurs dans le domaine de l'analyse globale des écritures manuscrites médiévales, nous



avons expertisé des descripteurs qui vérifient un certain nombre de conditions. Pour vérifier que l'on caractérise l'écriture et non le contenu du texte lui-même ni le style personnel du scripteur, l'inclinaison du texte, sa mise en page ou sa taille, les descripteurs de formes doivent être robustes, donc doivent pouvoir se calculer sans segmentation préalable des images et doivent résister au *changement de luminosité* et de contraste et au changement de résolution. Les descripteurs doivent être également *invariants au scripteur*, aux *contenus des textes*, à la *taille de l'échantillon* de texte, au *changement d'échelle*, au *changement de ratio* et à la *rotation*.

### 4.3 Application de la cooccurrence sur les écritures médiévales

La cooccurrence a été largement utilisée comme moyen de caractériser une texture en analyse d'images. Les images de documents présentent aussi des textures par la répétition des motifs réguliers des caractères, des mots et des lignes de texte. Cependant nous ne voulons pas mesurer la mise en page ni décrire la gestion des espaces (densité des traits, interlignes...). Nous cherchons plutôt à caractériser les écritures. Nous allons utiliser la cooccurrence de façon à ne mesurer que les variations des formes elles-mêmes et non les variations des formes entre elles. Pour cela nous devons effectuer de très faibles déplacements et nous assurer que l'on ne compare pas verticalement une ligne de texte avec les lignes adjacentes ou recouvrir horizontalement une lettre avec les lettres voisines. Par conséquent nous avons calculé les cooccurrences sur des images qui ont été normalisées manuellement pour qu'elles présentent toutes un corps de texte de 30 pixels de hauteur et nous avons limité les déplacements à moins de la moitié de la taille du corps des lignes de texte (FIG. 8). La normalisation de l'échelle des images par rapport au corps des textes est aussi nécessaire pour ne pas influencer la comparaison des observations sur des tailles de texte trop différentes (contrairement aux paléographes qui travaillent sur des images à l'échelle 1:1) (FIG. 9).

Nous cherchons plutôt à caractériser les écritures donc nous avons utilisé la cooccurrence de façon à ne mesurer que les variations des formes elles-mêmes et non les variations des formes entre elles. Nous avons effectué de très faibles déplacements afin de garantir une indépendance par rapport aux lignes et aux mots.

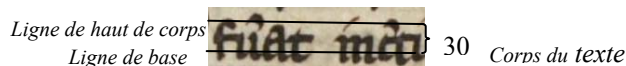


FIG. 8 - Normalisation du corps de texte

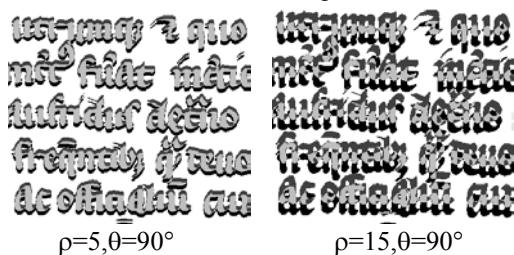


FIG. 9 - Calcul de matrices de cooccurrence sur une image binaire

La cooccurrence se généralise aux images en niveaux de gris, et donne des matrices de taille  $N_g \times N_g$  avec  $N_g$  le nombre de niveaux de gris de l'image pour chaque coordonnées trigonométrique  $(\rho, \theta)$ .

$$Co(\rho, \theta) = \frac{1}{N} \left[ \sum_{x,y} I(x,y) \cdot I(x+dx, y+dy) \right]_{i,j} = \frac{1}{N} [M_{i,j}]_{i,j=0..N_g-1}$$

(Avec  $I(x,y)$ : image d'origine,  $I(x+dx, y+dy)$ : image traduite et  $M_{i,j}$ : matrice de cooccurrence)

Nous avons choisi d'utiliser initialement un maximum d'information et de prendre un pas très fin pour les valeurs de  $\rho$  et de  $\theta$ . Nous avons utilisé 16 directions ( $\theta \in [0..15]$ ) et 15 déplacements possibles ( $\rho \in [1..15]$ ) soit  $16 \times 15$  matrices au maximum. Les valeurs des pixels ont été ramenées de 256 à 16 valeurs différentes. Une subdivision plus fine des valeurs de gris n'apporte pas d'information complémentaire pour des images de manuscrits qui sont constituées essentiellement de traits. Chaque écriture décrit une signature différente suivant les valeurs de  $\rho$  et  $\theta$  (FIG.10).

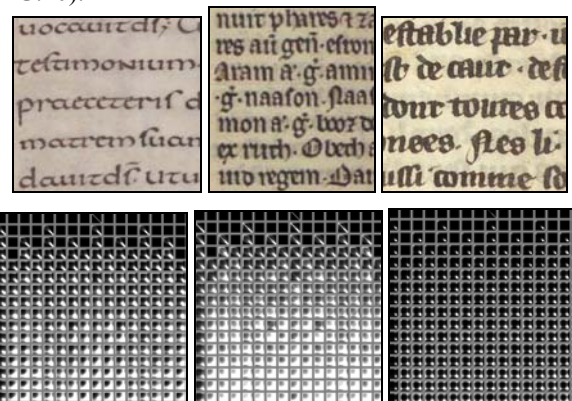


FIG. 10 - Matrices de cooccurrences relatives à quelques exemples de planches de styles différents

### 4.4 Analyse des descripteurs

Nous analysons les données des  $n$  observations décrites par  $p$  variables avec  $p$  égal au nombre de matrices de cooccurrence non nulles suivant  $\rho$  et  $\theta$ , multiplié par un nombre de descripteurs issus des travaux de Haralick [HAR 73] que nous présentons ci-dessous :

1	Second Moment Angulaire ou énergie	$f_1 = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P(i,j)^2$	Cette caractéristique mesure l'homogénéité, et détecte le degré de dispersion d'une texture.
2	Moment de la Différence de l'élément	$f_2 = \sum_{k=0}^{N_g-1} k^2 \times P_{x-y}(k)$	Cet indice mesure le degré de contraste ou de variation locale présent dans une image.
3	Corrélation	$f_3 = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-\mu_x)(j-\mu_y) P(i,j)}{\sigma_x \sigma_y}$	Cet indice décrit les corrélations entre les lignes et les colonnes de la matrice de cooccurrence et mesure de ce fait les dépendances linéaires entre les niveaux de gris dans une image.
4	Variance	$f_4 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-j)^2 P(i,j)$	Une forte valeur caractérise une texture fine.
5	Moment de la Différence Inverse	$f_5 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{P(i,j)}{1+ i-j }$	Une forte valeur indique que les éléments texturaux sont de grande taille.
6	Moyenne des sommes	$f_6 = \sum_{k=0}^{k=2N_g-1} k \times P_{x+y}(k)$	Moyenne des projections $P_{x+y}$

7	Variance des sommes	$f_7 = \sum_{k=0}^{k-2N_g-1} (k-f_0)^2 \times P_{x-y}(k)$	Variance des projections $P_{x-y}$
8	Entropie des sommes	$f_8 = - \sum_{k=0}^{k-2N_g-1} P_{x-y}(k) \times \log(P_{x-y}(k))$	Entropie de $P_{x-y}$
9	Entropie	$f_9 = - \sum_{i=0}^{i-2N_g} \sum_{j=0}^{j-2N_g} P(i,j) \times \log(P(i,j))$	C'est un indicateur de désordre.
10	Variance Différences	$f_{10} = \sum_{k=0}^{k-2N_g} (k-m_{x-y})^2 \times P_{x-y}(k)$	Variance de $P_{x-y}$ avec $m_{x-y} = \sum_{k=0}^{k-2N_g} k \times P_{x-y}(k)$
11	Entropie Différences	$f_{11} = \sum_{k=0}^{k-2N_g} P_{x-y}(k) \times \log(P_{x-y}(k))$	Entropie de $P_{x-y}$
12	Entropy Measure	$f_{12} = \frac{HXY - HXY1}{\max\{HX, HY\}}$ $HXY = f_9$	$HX = - \sum_{i=0}^{i-2N_g} P_X(i) \times \log(P_X(i))$ $HY = - \sum_{j=0}^{j-2N_g} P_Y(j) \times \log(P_Y(j))$ $HXY1 = - \sum_{i=0}^{i-2N_g} \sum_{j=0}^{j-2N_g} P(i,j) \times \log(P(i,j) \times P_Y(j))$

Avec

$$P_X(i) = \sum_{j=0}^{j-2N_g} P(i,j) \quad P_Y(j) = \sum_{i=0}^{i-2N_g} P(i,j) \quad P_{x-y}(k) = \sum_{i=0}^{i-2N_g} \sum_{j=0}^{j-2N_g} P(i,j)$$

$P_{x-y}(k)$  sont respectivement les projections

selon les axes horizontal, vertical et oblique de la matrice normalisée  $P$ .

Nous obtenons donc  $n$  points dans  $\mathbb{R}^p$  avec  $p=216 \times 12$ ,  $n$  étant le nombre d'images d'écritures observées. L'espace des caractéristiques est bien trop grand par rapport au nombre d'observations  $n$  pour un classifieur. Nous pensons qu'il existe parmi les  $p=2592$  variables, un nombre limité de facteurs qui peuvent faire apparaître les classes d'écritures. Un travail manuel de sélection des caractéristiques serait trop long et fastidieux. Il est donc nécessaire de réduire le nombre de descripteurs par une analyse statistique de la variance. Cette analyse nous a permis de trouver les descripteurs corrélés et donner un nombre réduit de facteurs qui sont des combinaisons linéaires des  $p$  variables d'origine. L'analyse des données est encore l'occasion de mener une analyse canonique de la proximité des classes puis comparer les résultats avec ceux des experts.

## 5 Analyse des résultats

L'analyse discriminante (AD) [DUD 00] a permis de trouver les vecteurs permettant de projeter toutes les observations définies dans un espace à  $p$  dimensions, sur un plan qui discrimine une grande majorité de classes (FIG.11). Le fait d'obtenir une majorité de classes séparées signifie qu'il existe des combinaisons linéaires de descripteurs qui peuvent résoudre notre problème de discrimination des écritures médiévales. Nous avons obtenu une bonne dispersion des classes : 1. *Caroline*, 3. *Cursiva Libraria*, 4. *Cursiva Formata*, 5. *Cursiva Currens*, 8. *Textualis prescissa*, 9. *Textualis Quadrata*, 10. *Textualis Semi-Quadrata*, 12. *Textualis Formata*, 13. *Textualis Libraria* et 14. *Textualis Currens*. La matrice de confusion a donné des taux de discrimination assez satisfaisants (de 48% pour la classe 12. *Textualis Formata* à 100% pour la classe 5. *Cursiva Formata*) pour les types d'écritures relatifs à ces classes. Les exceptions concernent les classes 2. *Gothique* et 7. *Textualis* non considérées comme de vraies familles ainsi que la 8. *Textualis Prescissa* et la 14. *Textualis Currens* qui ne sont pas significatives vu leurs nombres très réduits de représentants.

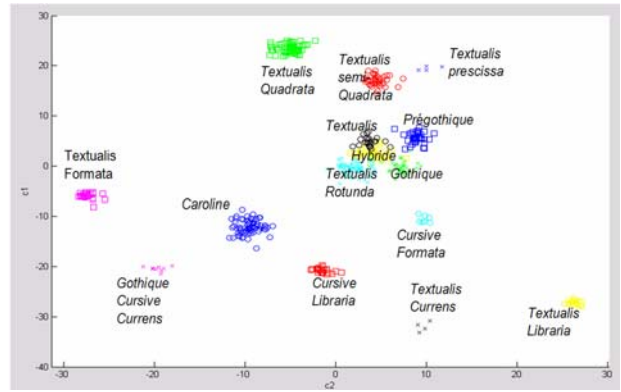


FIG. 11 - Résultats de l'AD pour les 15 classes

Quant aux 2. *Gothique*, 6. *Hybrida*, 7. *Textualis*, 11. *Textualis Rotunda* et 15. *Prègothique*, elles sont les moins bien séparées par l'AD et cela a entraîné des taux de confusions assez importants entre ces classes. Nos résultats montrent qu'il existe de véritables classes paléographiques au sens de la reconnaissance des formes. Certaines classes sont nettement séparées et forment de véritables familles bien identifiées. Ce sont les familles des écritures les plus détaillées (8. *Textualis Prescissa*, 9. *Textualis Quadrata*, 10. *Textualis Semi-Quadrata*, 12. *Textualis Formata*, 13. *Textualis Libraria*, 14. *Textualis Currens*) et les familles 1. *Caroline*, 3. *Gothique cursiva*; elles sont nettement séparées des autres classes. Les quatre classes confuses qui sont la 2. *Gothique*, la 7. *Textualis*, la 15. *Prègothique* et la 6. *Hybrida* ne constituent pas de véritables classes d'écritures homogènes au sens de l'analyse d'images. Les classes 2. *Gothique* et 7. *Textualis* contiennent des écritures non suffisamment renseignées par les paléographes et qu'il est donc normal que ces classes génériques soient confuses avec les sous-familles respectives. Les écritures prégothiques sont des écritures transitoires entre les écritures *carolines* et *gothiques* et ne peuvent pas être séparées par une analyse discriminante. En omettant les classes confuses les plus problématiques qui sont la 2. *Gothique*, la 7. *Textualis*, la 15. *Prègothique* et la 6. *Hybrida*, nous obtenons 11 classes correctement séparées

Le taux moyen de discrimination est passé de 59% à 81%. Il pourrait s'améliorer si nous disposions d'effectifs plus équilibrés et d'une meilleure représentation des classes 8. *Textualis Prescissa* et 14. *Textualis Currens*. Les 2 classes de faibles effectifs à savoir 8. *Textualis Prescissa* et 14. *Textualis Currens* ne peuvent pas être analysées du point de vue statistique, ce qui explique leurs faibles taux de discrimination.

En conclusion, il existe bel et bien des classes d'écritures qui paraissent compatibles avec l'expertise paléographique (selon la Bibliothèque de Grande Bretagne [BGB]) et la cooccurrence constitue une bonne mesure pour différencier les différentes écritures.

Tableau - Matrice de confusion obtenue par analyse discriminante sur 11 classes en utilisant les 12 caractéristiques de Haralick (f1à f12)

1. Caroline	2. Cursiva Libraria	3. Cursiva Formata	4. Cursiva Currens	5. Textualis Prescissa	6. Textualis Quadrata	7. Textualis Semi-Quadrata	8. Textualis Rotunda	9. Textualis Formata	10. Textualis Libraria	11. Textualis Currens	Total des Classes
91%	82%	71%	89%	50%	88%	63%	85%	70%	70%	50%	81%

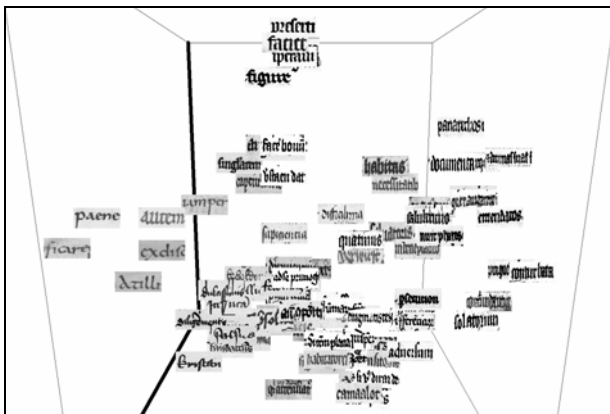


Fig. 12 - Affichage de quelques exemples de mots par classe d'après l'AD

## 6 Conclusion et perspectives

Après avoir établi le cadre de notre travail par rapport à la science de la paléographie, nous avons posé le problème de la classification des types d'écritures. Nous avons défini une approche globale comparant directement le type d'écriture à partir de zones de texte quelconques dans des documents. Nous avons choisi de travailler avec les indices statistiques de Haralick pour décrire nos matrices de cooccurrence afin d'avoir un nombre plus réduit de descripteurs par image.

Après une décorrélation des données par une analyse factorielle, nous avons constaté que nos descripteurs d'images basés sur des mesures statistiques de cooccurrence permettent de retrouver approximativement les classes d'écritures définies par la Bibliothèque de Grande Bretagne [BGB]. Les résultats de l'analyse discriminante ont été concluants et nous ont permis d'avoir dans plusieurs cas des séparations correctes des classes d'écritures. Nous avons obtenu un taux de 81% de discrimination global lorsque nous avons éliminé les quatre classes posant des problèmes de sous-représentation statistique ou bien d'absence de précision.

Le passage d'une famille à une autre n'étant jamais brusque et certaines écritures peuvent présenter un mélange de caractéristiques des écritures qui ont contribué à leurs formations, nous devons remplacer pour ces écritures l'analyse discriminante par une analyse qui mesure le taux de mélange avec les autres classes bien définies.

## Références

- [DER 03] Derolez A., "The Palaeography of Gothic Manuscript Books", from the Twelfth to the Early Sixteenth Century", Cambridge Studies in Palaeography and Codicology, Cambridge University Press, 2003.
- [CRE 95] Crettez J. P., "A set of handwriting families : style recognition", International conference on Document Analysis and Recognition, Vol 1, pp 489, 1995.
- [VIN 00] Vincent N., Boulétreau V., Sabourin R., Emptoz H., "How to use fractal dimensions to qualify writings and writers, Fractals", World Scientific, Vol 8, n°1, pp.85-97, 2000.
- [EGL 98] Eglin V., "Contributions à la structuration fonctionnelle des documents imprimés. Exploitation de la dynamique du regard dans le repérage de l'information", Thèse de Doctorat, INSA de Lyon, 1998.
- [MOA 02] Moalla I., Alimi A.M. Ben Hamadou A., "Extraction of Arabic text from multilingual documents", IEEE International Conference on Systems, Man and Cybernetics, Tunisia, 2002.
- [TAN 98] Tan T. N., "Rotation Invariant Texture Features and Their Use in Automatic Script Identification", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 7, pp. 751-756, 1998.
- [WOO 95] Wood S. L., Yao X., Krishnamurthi K., Dang L., "Language Identification for Printed Text Independent of Segmentation", Proc. IEEE ICIP, pp. 428-431, 1995.
- [AIO 99] Aiulli., Simi M., Sona D., Sperduti A., Starita A., Zaccagnini G. SPI: a System for Palaeographic Inspections. AIIA Notizie, vol. 4, pp 34-38, 1999.
- [CIU 05] Ciula A., "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis", Digital Medievalist 1.1, 2005.
- [HAR 73] Haralick R. M., Shanmugam K. and Its'Hak Dinstein. "Textural Features for Image Classification", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 3, pp. 610-621, 1973.
- [DUD 00] R. O. Duda, P.E. Hart, D.G. Stork, "Pattern classification", second edition.
- [YOS04] Yosef I. B., Kedem K., Dinstein I., Beit-Arie M., Engel E., "Classification of Hebrew Calligraphic Handwriting Styles: Preliminary Results", DIAL 2004, pp 299-305.
- [LIK 90] Likforman S. L., Maitre H., Sirat C. "An expert vision system for analysis of Hebrew characters and authentication of manuscripts", *Pattern Recognition*, Vol. 24, no 2, 1990, pp. 121-137
- [BV] <http://www.villevallenciennes.fr/bib/fondsvirtuels/microfilms/accueil.asp#item>
- [BGB] <http://prodigi.bl.uk/illcat/searchMSNo.asp>.