Contribution to the Discrimination of the Medieval Manuscript Texts: Application in the Palaeography

Ikram Moalla[1,2], Frank LeBourgeois[2], Hubert Emptoz[2], and Adel M. Alimi[1]

[1] REsearch Group on Intelligent Machines (REGIM),
University of Sfax, ENIS, DGE,
BP. W-3038 - Sfax – Tunisia
{ikram.moalla, adel.alimi}@ieee.org
[2] Laboratoire d'InfoRmatique en Images et Systèmes d'information (LIRIS),
INSA de Lyon-France
{Flebourg, Hemptoz}@rfv.insa-lyon.fr

Abstract. This work presents our first contribution to the discrimination of the medieval manuscript texts in order to assist the palaeographers to date the ancient manuscripts. Our method is based on the Spatial Grey-Level Dependence (SGLD) which measures the join probability between grey levels values of pixels for each displacement. We use the Haralick features to characterise the 15 medieval text styles. The achieved discrimination results are between 50% and 81%, which is encouraging.

1 Introduction

The Document Image Analysis is a particular research domain which is situated between images analysis, pattern recognition and human sciences especially the science that studies the history of texts. At present time, this research domain is spreading with the succession of the digitization of the ancient manuscripts of the cultural heritage notably in libraries and national archives *etc.* This revolution stimulates new research domains like the automatic extraction of the information for a better accessibility and a correct indexing of digitized documents. Among metadata which can be extracted, the writings styles brings additional information to the contents of the texts. The text layout represents a piece of information introduced in consciously or unconsciously by the writer which can be used to date, authenticate or index a document. The layout of a printed document is characterised by its physical structure and the characters typography (typestyle, size, font *etc.*) while the presentation of an ancient manuscript conceals other levels of interpretation such as the author's personal style of writing, the used calligraphy and the appearance of the document. The philology is a research field which study ancient languages, their grammars, the history and the phonetics of the words in order to educate and understand ancient texts. The philology is mainly based on the content of texts and concerns handwriting texts as well as printed documents. The paleography is a complementary discipline of the philology which collects handwritten texts corpus and knowledge accumulated on these documents. The paleography studies the layout of old manuscripts and their evolutions whereas the classic philology studies the

content of the texts, the languages and their evolutions. The goals of the palaeographic science are mainly the study of the correct decoding of the old writings and the study of the history of the transmission of the ancient texts. The palaeography is also the study of the writing style, independently from the author personal writing style, which can help to date and/or to transcribe ancient manuscripts. The target of this work consists of making a first methodological and applicable contribution to the automatic analysis of writing styles of old manuscripts for the service of the research in history of texts and for the palaeography science. We are interested more in ancient Latin manuscripts of the Middle Ages which precedes the Renaissance period before the emerging of the printing. The definition of the style is multiple and complicated. We are going to concentrate on a visual and perceptive approach of the style of writings which can be studied with images analysis tools. The recognition of the handwriting style which is connected to the historical period and/or the geographical localization independently of the personal style of the writer constitutes the main problem of our work.

2   The History of the Latin Writings

We present briefly the various Latin writings and their evolutions in Europe. Since the end of I^{est} century before J.-C, writings were transformed according to the usages. Since the $VIII^{th}$ until XIIth century, the Caroline was wide spread in the West.
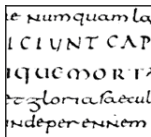
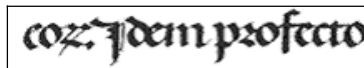

Fig. 1. Caroline sample

Fig. 2. Gothic sample

It evolved towards jagged forms to give birth in England to the Gothic writing, which spread in all the Northern Europe.

At the end of the XIVth century, the first humanists resumed the Caroline and created the humanistic. It was that writing which was adopted for printing and which became the basis of our modern writings. For palaeographers, the change from a writing to an other was not made in a radical way but by a slow and progressive evolution, which explains that it is difficult to identify categorically a given writing. For example we observe texts written in Caroline style which contain elements of the Gothic writing. Thus, the palaeographer should be able to quantify exactly the part of mixture of the writings families. For example the class of *Protogothic* writing is an intermediate writing style between the Caroline writing and the Gothic writing (Figs. 1, 2).

Since the XIIth century, the number of observed writing styles in Europe has exceptionally increased. Consequently, the work of palaeographers becomes more difficult especially with the evolution of the *Caroline* into *Gothic* (Fig. 3), and the division of *Gothic* into sub-families such as *Cursive Gothic* scripts, *Textualis Gothic* etc. Like the evolution of the *Caroline* into *Gothic*, the evolution into *Cursive Gothic* script then into *Batarde Gothic* thereafter into *Textualis Gothic* has been gradually made (Fig. 4).
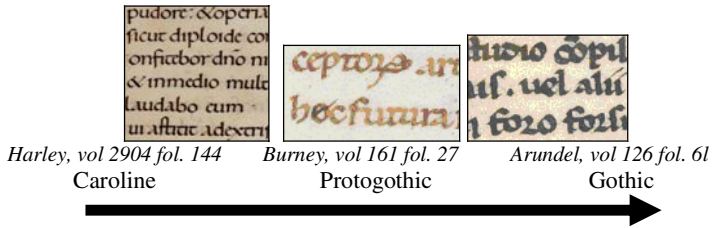
Harley, vol 2904 fol. 144   *Burney, vol 161 fol. 27*   *Arundel, vol 126 fol. 6l*
Caroline   Protogothic   Gothic

Fig. 3. Progressive evolution of the Caroline into Protogothic then into Gothic [BL]

Arundel vol 85 fol 1   Arundel vol 249 fol 5   Burney vol 335 fol 200
Gothic   batarde   Textualis

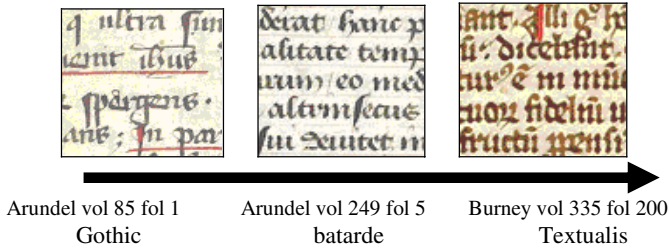**Fig. 4.** Samples of the evolution from cursive Gothic script into batarde Gothic then into Textualis Gothic [BL]

ms Thott vol 5554 fol 189v   *ms vol 131 fol 86r*   *ms vol 80 fol 163v*
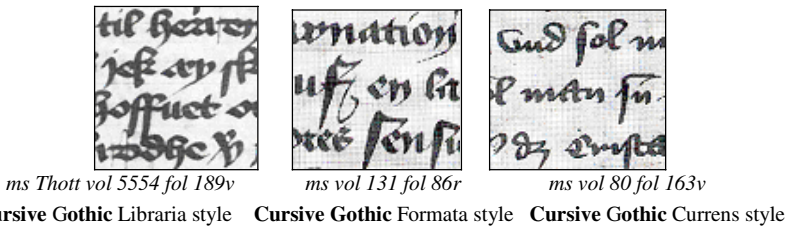**Cursive Gothic** Libraria style   **Cursive Gothic** Formata style   **Cursive Gothic** Currens style

Fig. 5. Samples of text images representing three sub families of cursive script between the 8th and the 16th century [1]

Yates Thompson   *Arundel Psalms*   *La bible*   *Burney vol 333*
Textualis Gotic Quadrata style   **Textualis Gotic** Semi-Quadrata style   **Textualis Gotic** Prescissa style   **Textualis Gotic**
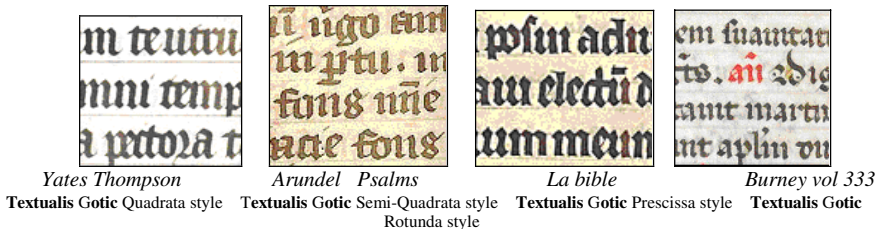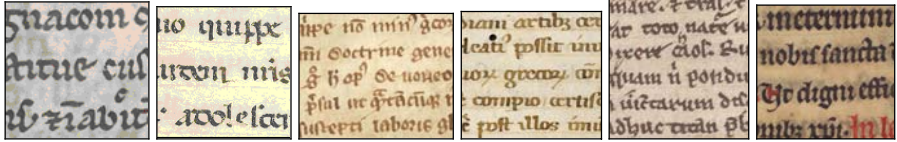   Rotunda style

Fig. 6. Samples of texts images representing the Textualis sub-families of styles between the 8th and the 16th century

The diversification of the writing families in Europe increased until the Renaissance and witnessed the development of writing subfamilies inside every big Gothic family.

Arundel vol 159 fol 5   Burney vol 239 fol 1   Burney vol 236 fol 2   Burney vol 235 fol 4   Burney vol 224 fol 3   Harley 928 fol 30

Fig. 7. Samples of texts images representing the big variation intra-classes for an example of Textualis Gothic Rotunda style [BL]

So we can distinguish several *Cursive Gothic* subfamilies of (Libraria, Formata and Currens) shown in Fig. 5. Also, the Fig. 6 shows several subfamilies of *Gothic Textualis* such as the *Quadatra*, the *Semi-Quadrata,* the *Prescissa*, the *Rotunda*, *etc.*

Fig. 7 shows the variability of writings inside the same sub-family as for the *Textualis Gothic Rotunda* class. It illustrates the difficulty in terms of image analysis to define the right features that describes the writing styles in order to find the homogeneity between the various samples of the same writing.

3   State of Art

We can find several work on the characterization of writings for different applications like the checking and the authentification of writer, the pre-classification of writings in terms of legibility for a better recognition in the automatic sorting of the mails and checks. All these studies are related to our problem but these contributions are not all directly re-exploitable for the paleographic study. The distribution of images directions was used to identify the different writings style for their recognition [2]. Fractal analysis measures the degree of autosimilarities in an image; it is a good measure of a writer's style that can serve to classify writings according to their legibility and to detect a modification of a writer for the early diagnostic of Alzheimer's illness [3]. Fractal indication is also susceptible to characterize the different alphabets in the printed texts. [4] characterized different text styles using complexity measures from shapes, legibility and compactness independently of the used alphabet. We can refer other works susceptible to be reused for the recognition of medieval writing such as the recognition of scripts (of words in a particular alphabet) in the multilingual documents. These works use the similarity of graphemes [5], the texture [6], or the analysis of projection profile [7] etc.

The System for Paleographic Inspections (SPI) [8], represents the only tentative for the realization of an automatic assistance system in paleography. [9], it is a local approach that tries to replicate the work of the paleographers. The method consists of isolating manually the representative characters of a writing and to compare them to referential characters from a paleographic database labeled manually. The comparison uses the tangent distance and the rule of the *k* nearest neighbor (*knn*) that gives *k* characters the nearest references to the new character. The system SPI only used for testing 37 documents and 4 images per styles and some images are descended from the same documents which is neither representative nor sufficient.

4 Our Proposition

We suggest to recognize the writing styles by using new image analysis methods to assist the historians in the classification and the dating of old Latin manuscript. Indeed, every historical period has been characterized by one or several types of writings. Therefore, the recognition of documents writings allows to know its date and/or its geographical origin.

We are not going to study the page layout of texts, the density of writings, the overlapping of characters, the concentration of diacritics which represent much susceptible information to characterize the style of a document. We limit our work to the classification of the writings into categories defined by paleographers.

Our domain of studies covers the old Latin writings of the VIIIth century until the XVIth. The study of Latin writing preceding the VIIIth century such as the Oncial or the cursive writing doesn't interests the paleographers. By contrast, the assistance to the medieval writing expertise is very useful since the XIIth century. It is for differentiating between main writing families (*Caroline* and *Gothic*) then to finely classify them into subfamilies (*Protogothic*, *Cursive Gothic*, *Hybrid Gothic* and *Textualis Gothic*) and then into more precise subgroup (*Rotunda*, *Quadrata*, *Semi-Quadrata*, *Prescissa*, *Libraria*, *Currens* and *Formata*) for the *Textualis Gothic* and (*Libraria*, *Formata* and *Currens*) for the *Gothic cursive* (see Fig. 7).

Our work focuses on the extraction of sufficiently discriminative features in order to be able to differentiate the biggest number of possible Latin writings. This study allows checking the feasibility of an automatic images analysis system that helps paleographers. First we examined the distances between the classes for studying the consistency between results of the images analysis and paleographic expertise.

Second, we refine the discrimination between the main Latin medieval writings then between the writings subfamilies as described in figure 6.
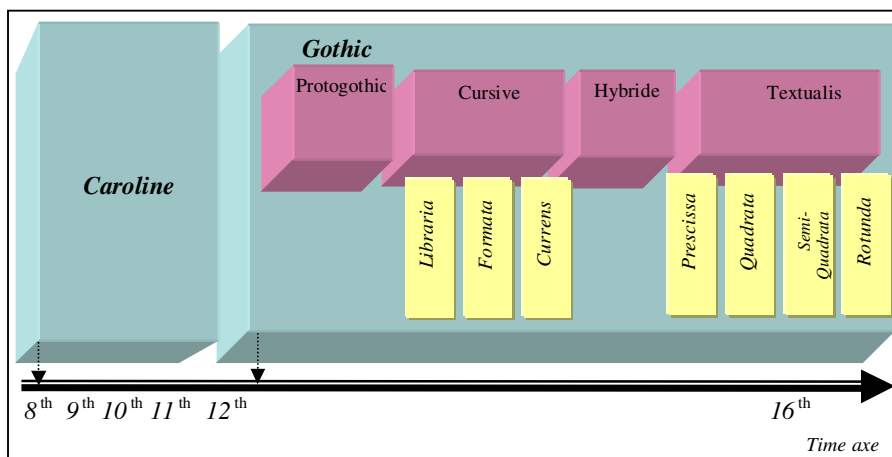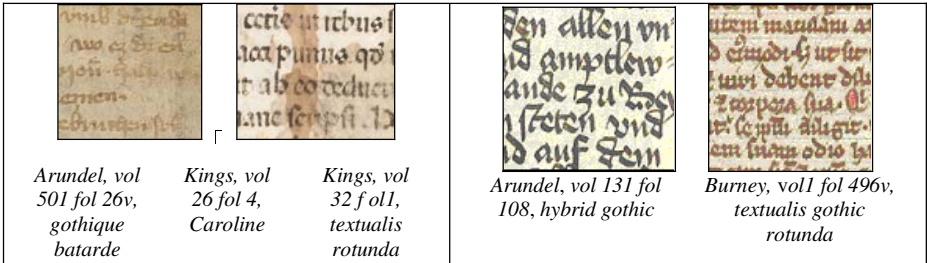

Fig. 8. Different subfamilies distribution of Latin style between the 8th and the 16th century

4.1 The Difficult Conditions

The development of a helping system for the old manuscript expertise is considered a difficult task for many factors:
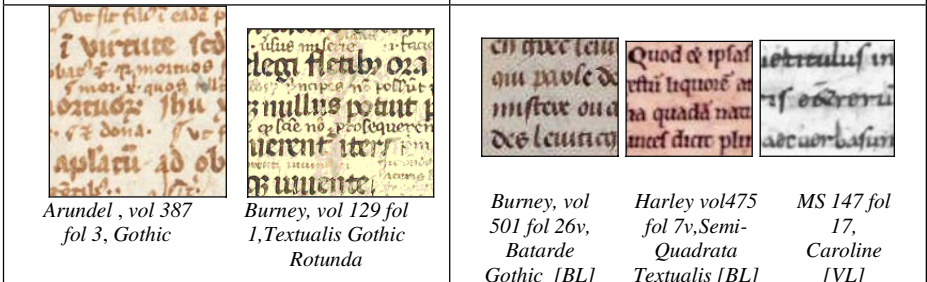
- The complexity of shapes of writings (Fig. 4, 5, 6), and the variability of writings from the same writing family (Fig. 8).
- The existence of hybrid writings that comes from a mixture of several writings (Fig. 1-3).
- The bad quality for manuscript conservation, for example the fading out of supports and inks (Fig. 9),
- The overlapping of lines and words (Fig. 10), and the presence of writing in the margin and/or between lines (Fig. 11).
- The bad quality of image origins; some colored images quality becomes deteriorated because of the digitization; and others from the digitization of books or microfilms in gray levels. Most images contain deteriorated areas due to a very strong compression (JPEG). Our samples are digitized with different resolutions (Fig. 12).

Therefore, within this difficult context, we analyze the image directly in gray levels without previous filtering, restoration or geometric correction. This choice deprives us from using a big part of the reusable works and in particular all those based on the segmentation.

Arundel, vol 501 fol 26v, gothique batarde   *Kings, vol 26 fol 4, Caroline*   *Kings, vol 32 f ol1, textualis rotunda*

Fig. 9. Samples representing two cases of ink fading out resulting a deterioration of characters [BL]

Arundel, vol 131 fol 108, hybrid gothic   *Burney, vol1 fol 496v, textualis gothic rotunda*

Fig. 10. Samples representing two cases of words and lines overlapping of [BL]

Arundel , vol 387 fol 3, Gothic   *Burney, vol 129 fol 1,Textualis Gothic Rotunda*

Fig. 11. Samples representing two cases of writing in margin and/or between lines [BL]

Burney, vol 501 fol 26v, Batarde Gothic [BL]   *Harley vol475 fol 7v,Semi-Quadrata Textualis [BL]*   *MS 147 fol 17, Caroline [VL]*

Fig. 12. Samples representing deteriorations due to a bad resolution

4.2 Our Approach

We distinguish two complementary approaches:

- **Local approach:** we try to replicate the work of paleographers, while attempting to establish some visual similarities between writings relied on very particular features of a letters writing (examples: 'r', 's', 'e', 'a'). Indeed, some particular letters are used by paleographers for the recognition of a writing. These letters must be taken inside words because their graphics change according to the writer when they are situated at the beginning or at the end of words [8] [9].
- **Global approach:** we do not try to replicate the work of paleographers, but to use a more suitable method for the automatic images analysis. The approach consists of analyzing statistically the whole image of a manuscript and to find features which describe writings. The global approach should guarantee the independency of the global measures from the text content, the writer's personal style, the used language, the used letters and of their frequency. If the sample size is meaningful, all the letters are represented and in particular the characteristic letters used by palaeographers.

Moreover, a global analysis allows the inclusion of some ornaments without affecting a statistical analysis because the text occupies a sufficient area.

The global approach advantages are very precious for the analysis of a great variety of documents having different qualities and origins. So we have chosen to work with this approach to overcome the difficult conditions described before.

Because of the lack of previous works in the domain of the global analysis of the medieval manuscripts writings, we have to find `image features` that verify the following criteria:

− The robustness: image features can be calculated without any image segmentation or any prior processing.
− The writer invariance: the measures should be independent of the writer.
− The invariance to the size: image features must be invariant to the size of the text sample.
− The change of scale: A writing must be invariant to the scale factor, but some images features require to resize the image so as the scale of different writings are comparable.
− The change of ratio: It is the most current geometric transformation to adjust images to an electronic document. The ratio height/width of an image must not influence the final decision. Image features can work differently on images having different ratios. Therfore, we suggest that image maintain all the same ratio.
− The rotation: A writing must remain the same whatever the image orientation can be. In image analysis, describers must be invariant to the same rotation applied to all images.

We suggest to achieve a classification system of writings. If the writing family is found and/or its rate of mixture with other writings is determined, we can give more or less precise date of the document.

4.3   Application of the Cooccurrence on the Medieval Writings

The cooccurrence has been used as a means for characterizing a texture in image analysis. The images of documents present also textures by the repetition of the regular characters, the words and the lines of the text. However we want neither to measure the page layout nor to characterize the management of spaces (density of features, spacing…), we would rather try to characterize writings. We use the cooccurrence just to measure writing variations and not the variations of shapes between themselves. Therefore, we have to do very weak displacements and be assured that we do not compare two adjacent lines or cover a letter horizontally with the neighboring letters. Cooccurrence must be calculated on texts that are normalized in size and displacements must be limited to less than half of the size of the text lines body. We normalize all the images of our experimental database with the overage text body roughly equals 30 pixels to allow the displacements of a distance that exceeds 15 pixels as a maximum.
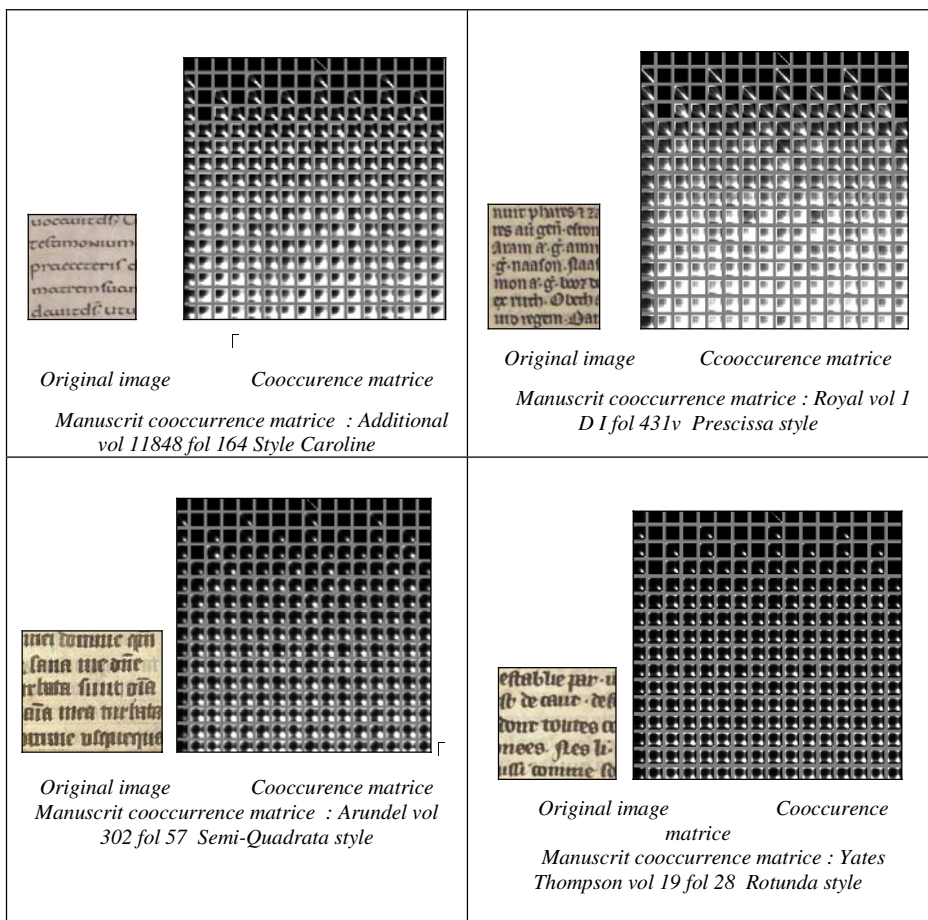

Original image   Cooccurence matrice

Manuscrit cooccurrence matrice : Additional vol 11848 fol 164 Style Caroline

Original image   Ccooccurence matrice

Manuscrit cooccurrence matrice : Royal vol 1 D I fol 431v Prescissa style

Original image   Cooccurence matrice
Manuscrit cooccurrence matrice : Arundel vol 302 fol 57 Semi-Quadrata style

Original image   Cooccurence matrice
Manuscrit cooccurrence matrice : Yates Thompson vol 19 fol 28 Rotunda style

Fig. 13. Cooccurrences matrices relative to some samples of different writings style

For each direction *theta (θ)* and displacement *rau (ρ)*, we have a cooccurrence matrix of size NgxNg with Ng is the number of gray levels of the image.

$$Coo(dx = \rho \cos \theta, dy = \rho \sin \theta) = \frac{1}{N}\left[\sum_{x,y} I(x, y) = i \cap I(x + dx, y + dy) = j\right]_{i,j} = \frac{1}{N}\left[M_{i,j}\right]_{(i,j)=0..Ng-1} \tag{1}$$

We use the maximum of information and take a very fine subdivision for the values of ρ and of θ. We have used 16 directions (θ∈[0..15]) and 15 displacements (ρ∈[1..15]) that is 16x15 matrices to the maximum. The values of pixels have been decreased from 256 up to 16 values. We do not keep matrices of cooccurrence for ρ=0, because they don't correspond to any displacement. The discreet nature of images does not permit to have more than 4 directions for the displacement of 1 pixel, 8 directions for a displacement of 2 pixels etc. It remains 216 non null matrices. Every writing is described by a different signature according to the values of ρ and θ (Fig. 13).

4.4 Verification of Criteria by the Cooccurrence Measures

The cooccurrence matrices relative to samples of different sizes of the same document are approximately similar. Information is considered incomplete for a very small size sample. If the image contains only some words, it does not exist enough information on the intermediate characters.

The cooccurrence is *invariant to text content* because the SGLD are similar on different text areas of the same document. The cooccurrence is *robust* because it does not require any image segmentation nor of text zones, lines, words nor of characters.

The image smoothing modifies greatly the SGLD for the small displacement of *rau* near 1. Because of the specific nature of digitized document, the image smoothing densifies the extreme values of matrices for (i,j)=(0,0),(0,15),(15,0) and (15,15).

The modification of the image ratio is equivalent to the calculation of the cooccurrence with a displacement ρ(θ) which describe an ellipse and not a circle. The impact on the cooccurrence matrix is equivalent to the change of scale but with a non constant displacement ρ. As we constitute a feature vector from the cooccurrence matrices in the order of ρ and θ, the rotation, the scale and the ratio modify the data position in the feature vector but not the information itself.

The cooccurrence preserves the same information about shapes after the main geometric transformations. But this information is not preserved anymore by the same matrices following ρ and θ. To guarantee that we compare the same information, it is necessary that the images have the same orientation, scale and ratio.

4.5 Images Features

We analyze *n* observations data described by *p* variable with *p* equal to the number of cooccurrence matrices non null multiplied by the number of 12 Haralick describers [10] (With a quantification into 16 values, for ρ and θ, the cooccurrence represents 216 non null matrices of 16x16 values). So we have *n* points in IR^p with p=216x12, *n* is the number of observed images writing.

The features' space are too big in relation with the number of **n** observations for a classifier. There is a limited number of factors among the **p**=2592 variables that participates in the categorization of writings. A manual work of features' selection would be too long and exhausting. Therefore it is necessary to reduce the number of describers by a statistical analysis of the variance.

This analysis allowed us to find the correlated variables and to give a reduced number of factors that are of linear combinations of the origin variable *p*. The data analysis leads to a canonical analysis of the class proximity, then to comparison of the results with those of experts.

5   Analysis of Results

Considering the following references for the 15 classes of the Latin writing:

1 : Caroline	6 : Hybride batard	11 : Textualis Rotunda
2 : Gothic	7 : Textualis	12 : Textualis Formata
3 : Cursive Libraria	8 : Textualis Prescissa	13 : Textualis Libraria
4 : Cursive Formata	9 : Textualis Quadrata	14 : Textualis Currens
5 : Cursive Currens	10 :Textualis Semi-Quadrata	15 : Protogothic

In order to have a general view of the 15 classes, we applied a global discrimination strategy onto the 15 classes.

While applying the PCA (Principle Component Analysis) with only one measure like *f10* that represents the variance of P_{x-y}, we get the factorial map of the Fig.14. This map represents 97% of the variance explained by the first two axes which proves that data is correlated and that we can reduce the number of the characteristics without losing information. This map shows that the different writting form clusters correspond approximately to the classes defined by the palaeographers. This "blind" analysis, without taking into account the classes, shows that the paleographers' classification is coherent and that the writings of the same class are near.

The cooccurrence constitutes a good measure to differentiate between the various writings. However, if these features explain well the variance of observations, they are not necessarily the most discriminative classes. Therefore, we are going to apply the discriminant analysis [11].
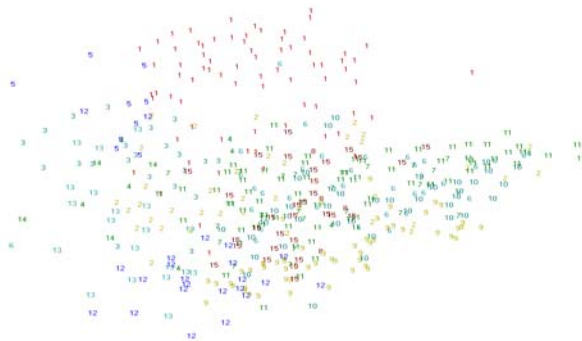
Fig. 14. PCA on the 15 classes with f10 characteristic

Contrary to the PCA, the Discriminant Analysis finds linear projections into a subspace that better discriminates a great number of classes if the features are relevant (Fig. 15). Getting a majority of classes separated proves the existence of linear combinations of describers which can solve the problem of medieval writing discrimination. We have obtained a good scattering of classes: 1. Caroline, 3. Cursive Libraria, 4. Cursive Formata, 5. Cursive Currens, 8. Textualis prescissa, 9. Textualis Quadrata, 10. Textualis Semi-Quadrata, 12. Textualis Formata, 13. Textualis Liraria and 14. Textualis Currens. The confusion matrix confirms the good results given by the satisfactory discriminating rates for the writing types relative to these classes (from 48% for the class 12. Textualis Formata up to 100% for the class 5. Cursive Formata). Exceptions concern classes 2. Gothic and 7. Textualis that are not considered as true families as well as the 8. Textualis Prescissa and the 14. Textualis Currens which are not enough statistically represented in our database.
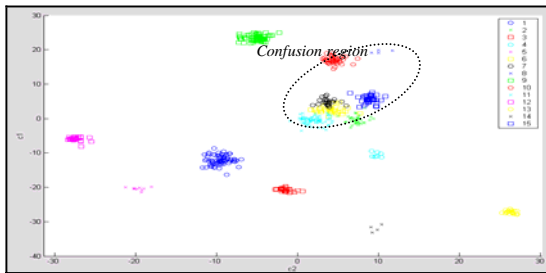

Fig. 15. Result of DA for 15 classes

The writing style *2. Gothic*, *6. Hybrid*, 7. *Textualis*, *11. Textualis Rotunda* and *15. Protogothic* are the least well separated by the discriminant analysis and show important confusion between them. The four confused classes that are the *2. Gothic*, the *7. Textualis*, the *15. Protogothic* and the *6. Batarde* do not constitute any real homogeneous writing classes from the image analysis point of view. We think that classes *2. Gothic* and *7. Textualis* contain writings non sufficiently described by paleographers and it is therefore normal that these generic classes are confused with their respective subfamilies. We think that *Protogothic* writings are transitory writings between *Caroline* and *Gothic* writings. Dendrogram analysis confirmed that the *Batarde* writing is a *hybrid* writing between the *Cursive Gothic* writings and the *Textualis Gothic* writings.When we omitted the most problematic classes which are the *2. Gothic*, the *7. Textualis*, the *15. Protogothic* and the *6. Batarde*, we obtained 11 correctly separated classes. Our results show that it exists coherence between image features and the palaeographic classes of medieval writings. We think that *Protogothic* writings do not constitute an independent class which cannot be discriminated from the *Caroline* and the *Gothic* writings. For the *Protogothic* writing, we can provide to palaeographers the rate of mixture between *Caroline* and *Gothic* by taking the distance from the centres of the respective classes. The average rate of discrimination moved from 59% to 81%. It can be improved if we will have a better equilibrated number of samples for classes 8. *Textualis Prescissas* and 14. *Textualis Currens*.

Table 1. Confusion matrix obtained by discriminative analysis onto 11 classes while using the 12 Haralick features

	1	2	3	4	5	6	7	8	9	10	11	%correct
1.Caroline	68	1	0	1	0	0	0	5	0	0	0	91%
2.Cursive Libraria	0	18	1	0	0	0	0	0	0	3	0	82%
3.Cursive Formata	0	1	5	0	0	1	0	0	0	0	0	71%
4.Cursive Currens	0	1	0	8	0	0	0	0	0	0	0	89%
5.Textualis Prescissa	0	0	0	0	2	0	0	2	0	0	0	50%
6.Textualis Quadrata	0	0	0	0	0	49	1	4	2	0	0	88%
7.Textualis Semi-Quadrata	0	0	0	0	0	4	26	9	2	0	0	63%
8.Textualis Rotunda	1	1	0	0	0	5	3	64	1	0	0	85%
9.Textualis Formata	0	1	0	1	0	2	0	1	16	2	0	70%
10.Textualis Libraria	0	2	1	1	0	0	0	0	1	14	1	70%
11.Textualis Currens	0	0	0	0	0	0	0	0	0	2	2	50%
Total	69	25	7	11	2	61	30	85	22	21	3	81%

6   Conclusion and Perspectives

We have exposed the problem of the classification of ancient manuscripts which is useful for the paleography science.

We defined a global approach which does not require the binarisation of images or text segmentation. We suggested analysing globally some text blocks which are enough representative of the writing style of the entire document. We chose to work with the cooccurrence and used the statistical features of Haralick to describe our matrices of cooccurrence in order to have a reduced number of image features.

Our images describers based on the statistical measures of cooccurrence allow to find approximately the classes of writings defined by the palaeographers after the decorrelation by a factorial analysis. The discriminant analysis provides a rate of 59% of global discrimination for the fifteen Latin classes. The discrimination rate increases up to 81 % when we eliminate the four classes causing problems which are not statistically well represented or because of absence of precisions. Indeed the proceeding from one family to another has never been abrupt and some writings can present a mixture of writings features that contributed to its formation. We mention the *Protogothic* and the *Hybrid* as examples. For these writings, we must replace the discriminant analysis by an analysis that measures the rate of mixture with the other definite classes. We also noticed that the *Gothic* and *Textualis* writings are only generic writings that have not been sufficiently described (a hypothesis that remain to be validated by experts in paleography). Contrary to character recognition or scripts separation, classification of medieval writings requires experts in Palaeography to valid our work and confirm the right classification of the images from our database. We found a lot of resources of images on the Web, but we are not sure that the classes given by paleographers are exact. We hope to get the help of paleographers to exploit a bigger number of these resources.

Moreover, we try to increase our collaboration with palaeographers in order to analyse the results from the image analysis point of view and to refine our approach to better fit their needs.

References

1. A. Derolez, "The Palaeography of Gothic Manuscript Books", from the Twelfth to the Early Sixteenth Century", Cambridge Studies in Palaeography and Codicology, Cambridge University Press, 2003. (http://www.moesbooks.com/cgi-bin/moe/39006.html).
2. J. P. Crettez, "A set of handwriting families : style recognition", International conference on Document Analysis and Recognition, Vol 1, page 489, Auguest 1995.
3. V. Bouletreau, "Vers une classification de l'écrit", Thèse de doctorat INSA de Lyon, 1997.
4. V. Eglin, "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, 13 Novembre 1998.
5. I. Moalla, A.M. Alimi and A. Ben Hamadou, "Extraction of Arabic text from multilingual documents", IEEE International Conference on Systems, Man and Cybernetics, Tunisie, Octobre 2002.
6. T.N. Tan, "Rotation Invariant Texture Features and Their Use in Automatic Script Identification", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 7, 1998, pp. 751-756.
7. S. L Wood, X. Yao, K. Krishnamurthi, L. Dang, "Language Identification for Printed Text Independent of Segmentation", Proc. IEEE ICIP'95, pp. 428-431, 1995.
8. Aiolli, F., M. Simi, D. Sona, A. Sperduti, A. Starita, and G. Zaccagnini. 1999. SPI: a System for Palaeographic Inspections. AIIA Notizie http://www.dsi.unifi.it/AIIA/ vol. 4: 34-38.
9. A. Ciula, "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis", Digital Medievalist 1.1, April 20, 2005
10. R. M. Haralick, "Statistical and structural approaches to texture", Proceedings of IEEE, vol. 67, no. 5, pp. 786{804, 1979.
11. R. O. Duda, P.E. Hart, D.G. Stork, "Pattern classification", second edition
    [VL] http://www.villevalenciennes.fr/bib/fondsvirtuels/microfilms/accueil.asp#item
    [BL] http://prodigi.bl.uk/illcat/searchMSNo.asp