Curvelets based Feature Extraction of handwritten shapes for ancient manuscripts classification

Guillaume Joutel, Véronique Eglin, Stéphane Bres, Hubert Emptoz

LIRIS – INSA de Lyon,
20 avenue Albert Einstein, 69621 Villeurbanne Cedex, France
{guillaume.joutel,veronique.eglin, stephane.bres, hubert.emptoz}@insa-lyon.fr

Abstract. The aim of this scientific work is to propose a suitable assistance tool for palaeographers and historians to help them in their intuitive and empirical work of identification of writing styles (for medieval handwritings) and authentication of writers (for humanistic manuscripts). We propose a global approach of writers' classification based on Curvelets based features in relation with two discriminative shapes properties, the *curvature* and the *orientation*. Those features are revealing of structural and directional micro-shapes and also of concavity that captures the finest variations in the contour. The Curvelets based analysis leads to the construction of a compact Log-polar signature for each writing. The relevance of the signature is quantified with a CBIR (content based image retrieval) system that compares request images and database images candidates. The main experimental results are very promising and show 78% of good retrieval (as precision) on the Middle-Ages database and 89% on the humanistic database.

Key-words: multiscale shape decomposition, curvelets, handwriting classification and retrieval, logpolar signature.

I. Introduction

In this project, we are interested in digitized Middle-Age (composed by copyists' texts from the 9th to the 15th century) and Humanistic manuscripts (essentially composed by authors' drafts from the 18th and 19th century) analysis for an expert classification and identification. The aim of this scientific work is to propose a suitable assistance tool for palaeographers and historians to help them in their intuitive and empirical work of *classification* of writing styles from different historical period for medieval manuscripts and *identification* of writers for humanistic and contemporary author's manuscripts. In those corpuses, writings styles bring important information on the writer which can be used to date, authenticate or index a document. The main problem consists in proposing robust solutions to the writings classification that must be independent of the author personal writing style (for the Middle-Ages manuscripts, see figure 1.1), or in opposite that must be strongly correlated to the personal style (for the author's manuscripts, see figure 1.2).

In this work, the efforts are not applied on the transcription of texts which is made by literary experts, but they are applied on the analysis of handwriting patterns in the image (bitmap) representations. The immediate objective consists in classifying and in characterizing copyist's and writers' contributions according to their own handwriting styles. The writers' classification is based on the definition of features that can be extracted over an entire textural handwriting sample or directly on the writing itself.

Document Recognition and Retrieval XIV, edited by Xiaofan Lin, Berrin A. Yanikoglu
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6500, 65000D, © 2007 SPIE-IS&T · 0277-786X/07/$15

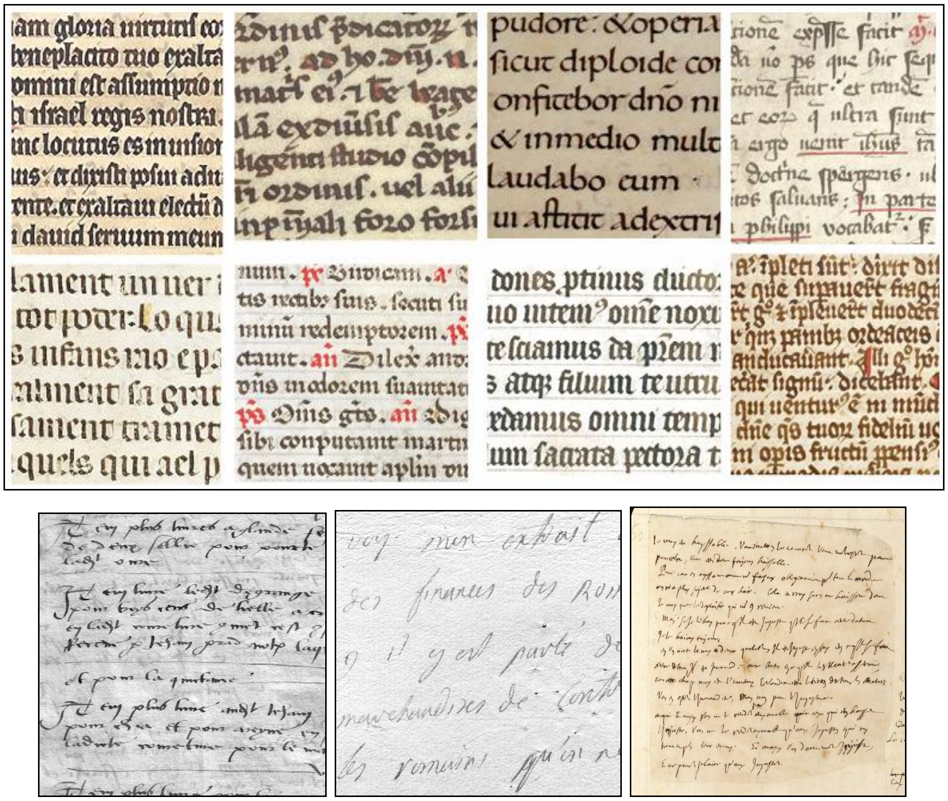


Fig. 1. 1.1 Examples of the great diversity of Middle-Ages (from the 11th to the 14th century), British Library. 1.2 And Humanistic manuscripts corpus (18th century), ENS Lyon Database.

Here, we propose a global approach of writers' classification based on Curvelets based features that are able to reveal typical morphological shapes properties: the *curvature* and the *orientation*. Those primitives are revealing of structural and directional micro-shapes and also of concavity that captures the finest variations in the contour. For palaeographers these primitives have a great interest for the manuscripts dating.

This work is currently used for Middle-Ages manuscripts classification (from the 9th to the 14th century) and for the retrieval of a particular author manuscript in Humanistic corpus. A typical example is given by the Montesquieu's corpus (philosopher of the 18th century). His manuscripts are characterized by a great diversity of writers: they were written by more than twenty secretaries having all different visual characteristics of writings. The pages of this collection can be considered as *drafts* with a lot of corrections, crossings out and scribbles. In both cases (Middle-Ages and Humanistics manuscripts) our purpose here is to prove that it is possible to characterize handwritings and to classify them into visual writers' families with the *orientation* and the *curvature* as single couple of discriminative feature.

II. Related works

As for now, studies for writing characterization are essentially based on standard handwriting patterns analysis in bi-level images, but no real method has been developed for the valorisation of this kind of specific historical corpus. General methods are either based on the consideration of local particular graphemes [3] or on a too macroscopic and general characterization that is not efficient for a relevant writers' authentication, [4]. In that context, we have chosen to propose a dual methodology based on two complementary approaches: a texture based approach that considers the handwriting in its global and homogeneous environment and a local shape based approach that considers the handwriting as a series of loops and right segments. There exist many discriminating features that allow matching different handwritings between two documents. The matching is generally performed between the same kinds of features which are measured on each document, [6]. But features, which capture the global characteristics of the writer's individual writing style are rarely associated to micro-features for finer details at the character level. The combination of a macro-feature analysis (that can also be regarded as texture-based analysis) and a local feature analysis is rarely proposed in the most recent researches. In this work we propose a unified approach which allow combining both of them for the writers' style classification and the individualities characterization.

Different distances are currently used to access the best matching between different set of handwriting samples. Discriminability of a pair of writing samples based on similarity value can be observed by studying their distributions when the pair arises from either the same writer or from different writers. Generally, in most writers' classification approaches, authors try to produce a set of exhaustive graphical characteristics which can efficiently describe all handwritings families. It is difficult to pretend to be exhaustive in such a description, so it is essential to work with expert users who are able to validate the measures that appear to be the most relevant. The signification degree that is assigned by the user could also guide the system to create a suitable distance between writing classes. In any cases, each computed measure should be evaluated in relation with all others. In this case, a Principal Component Analysis (PCA) considerably reduces the dimensionality of micro-features vectors, [10]. Within this kind of generic approach, it is possible to classify handwriting samples into visual distinct style classes. In our laboratory, we plenty use frequencies domain with Hermite and Gabor based features which reveal salient oriented graphemes on high frequencies contours at different scales and resolutions, [7].

The only methods that use wavelets decomposition for handwritings classification do not make any hypothesis on the information to measure, [8]. The general case is to compute several measures on wavelets coefficients but without any real link with the real meaning, like in [8] and then with [11]. In our proposition, we lead a Curvelets based analysis for a designed decomposition in curvature and orientation which have a real sense for paleographers. It is important to recall that those measures of orientation and curvature are obtained without any segmentation of the handwriting.

In handwriting classification which is our main goal, the most closely related works are [7] and [10]. In [7], authors propose an analysis of the variability of handwritings on the bases of two kinds of observations. The first kind of observations is a group of three classical measures which are semantically independent: thickness of the tracing, main body of the word and spatial density of characters. The second kind of observations is linked to the successive directions of the tracing. This work has been led on contemporary documents only.

The only work relative to the analysis of the Middle-ages documents referenced in [1] propose classifications procedures which are today debatable by palaeographers. Some recent studies have tried to develop more robust approach for a non supervised classification. In [12] for example, authors are interested in ancient Latin and Arabic manuscripts of the Middle-Ages before the Renaissance period and the emerging of the printing. They made the choice to analyze statistically the whole image of a manuscript and measure globally all patterns. This approach should guarantee the independency from the text content, the writer's personal style, the language used and letters frequencies. To do so, they use the cooccurrence approach which measures the relationship between intensity values. The cooccurrence is evaluated from the SGLD (*Spatial Gray-Level Dependence*) which is a join probability to observe the same intensity value between two different pixels according to their spatial relation. In characterization's term, the most closely work is [7]. In this work, they use the spectral domain of handwritten images by frequencies decomposition (Hermite transforms) and Gabor bank filtering for selective text information extraction. The handwriting characterization is done with a multi-scale analysis of emergent orientations. This could have been used for our first measure (orientation) but for curvature we should have searched for a third algorithm in addition to the two first already used. Our work differs from those one because it is dedicated to both Middle-Ages and contemporary handwritings styles. Moreover it uses a wavelets based tool, the Curvelets, which has never been used in this context.

III. Curvelets Transform for handwriting characterization

Middle-Ages and Humanistic manuscripts form two relevant historical sets which are well adapted to an automatic analysis. Historical handwritten documents can not be easily segmented in lines or in words. Existing methods have proved their limits with the increasing text non linearity, some unpredictable page layout, the irregularities of handwritten shapes, sometimes too small interline spaces and frequent words contacts (especially for Middle-Ages documents). The separation of text and non text areas becomes difficult in case of insufficient pen pressure (for Humanistic pages). Beside those difficulties, we can also mention the visual appearance of ancient documents backgrounds which are often characterized by additional noises that disturb the handwriting segmentation (see figure 1).

Traditional regular handwriting segmentation approaches have shown their inefficiency in ancient heterogeneous or degraded corpus. In that context, we can notice that a bi-level page segmentation generally damages the handwritings by merging words or text lines together. So as to bypass those drawbacks, we preferably will choose a threshold free methodology which can produce a selective mapping of the page image by capturing grey level textual informative areas. In that way, it is possible to carry out the study by considering visual emergent properties of the handwritten patterns as it is naturally made by the way of a human visual expertise. The second point, which must be revealed, deals with handwriting specificities. Indeed, the kind of writing device, the ink type and the manner they are employed can transform the global aspect of a curve and of a character. The lines thickness in a part of the handwriting must be considered as a variable value. Different tools, based on spectral domain of images or on perceptive approaches, could find here a real sense (fractal measures, Gabor and Hermite transforms, directional specialized wavelets coding…). But in our situation, the models which have shown the best compromise between a local and a global featuring are the wavelets. In this paper we are using a specialized version of directional wavelets: the Curvelets. The wavelets reveal isotropic elements at all scales and locations but do not describe well highly anisotropic elements and contain only a fixed number of directional elements independent of scale.

As presented in [5,13], the idea of curvelets is to represent a curve as a superposition of functions of various lengths and widths obeying the scaling parabolic law width $\cong$ length ^{2}.This can be done by decomposing the image into a series of disjoint scales. Each scale is then analyzed by means of a local ridgelet transform. So, Curvelet transform is based on multiscale ridgelet transform combined with a spatial bandpass filtering operation at different scales. It consists in a multiscale ridgelet which exists in a prescribed frequency band. The bandpass is set so that the curvelet length and width at fine scales are related by a scaling law width $\cong$ length ^{2}. With this ratio, the anisotropy increases with decreasing scale like a power law. With a dyadic decomposition of the frequency domain, the *length* of the localizing windows is doubled *at every other* dyadic subband, hence maintaining the fundamental property of the curvelet transform, see figure 2. The multiscale analysis decomposes the spectral domain into oriented sub-bands of frenquencies $f\in [0,1/2^{2S}]^2\backslash[0,1/2^{2S+1}]^2$, with $s\in N$ is the scale. This decomposition is well adapted to curve approximation.

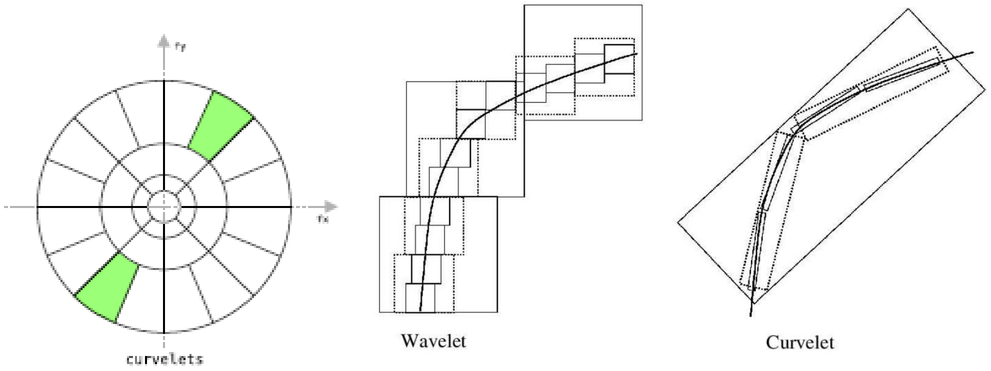


Fig.2. Dyadic decomposition of the spectral domain. Wavelet and a Curvelets decomposition obeying the scaling law width $\cong$ length ^{2}

Figure 3 illustrates the decomposition of the original image into frequential subbands (through the Radon transform) and followed by a spatial partitioning of each subband into blocks on which the ridgelet transform is finally applied. For all details concerning the theorical description of the curvelets, the reader can refer to [5].

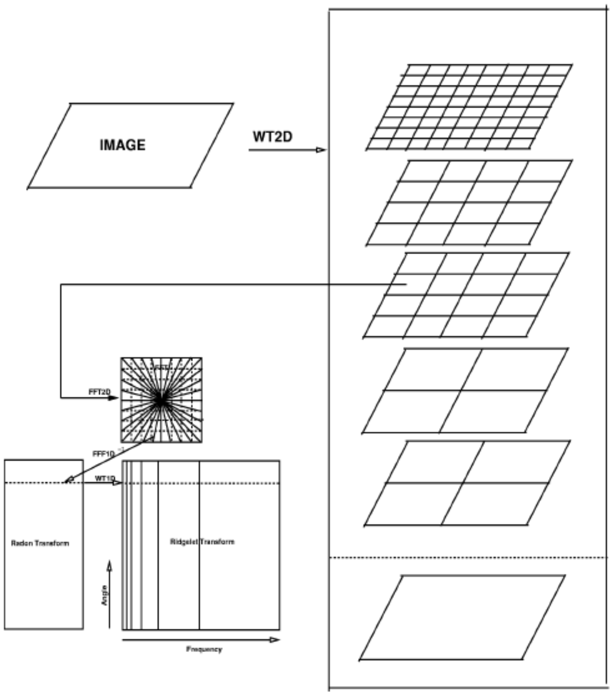


Fig.3. Digital Curvelet Transform flowgraph

Curvelets enjoy two unique mathematical properties, namely: first, the curved singularities can be well approximated with very few coefficients and in a non-adaptive manner, and then the Curvelets remain coherent waveforms under the action of the wave equation in a smooth medium. Practically, Curvelets transform is a new multi-scale geometrical transform in which units are indexed by their position, their scale and their orientation. The directionality concept is integrated and allows an optimal and compact representation of images that contain objects with clear frontiers as for the writings we are working on. In our case, one essential advantage of the Curvelets is their adaptability to different manuscripts corpus with very different writing styles and no common shapes properties (see figure 1).

Here, we have been interested by the ability of curvelets transform to reveal curvature and orientations of handwritten shapes: these are two complementary dimensions considered as fundamental features of handwritings by paleographer experts. The *curvature* at the main feature of contours description and the *orientation* is the main feature for feature for right segments quantization. Those primitives have been represented in figure 4 with the reconstruction of the handwriting sample by a curvelets coefficients quantization.

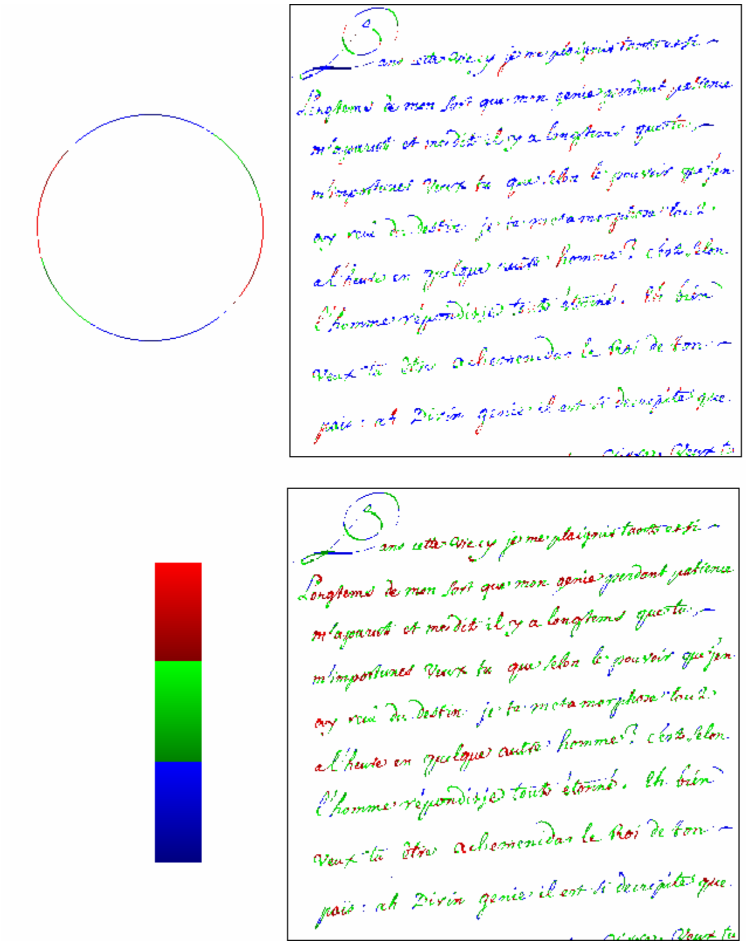


Fig.4. Reconstruction of a sample of Montesquieu's handwriting with the location of orientations (on the left) and the location of curvatures (on the right). The oriented circle calibrates the orientation and the coloured scale (from red to blue) indicates the intensity of curvature in the handwriting.

On the top part of the figure 4, one can see a representation of orientations extracted from a sample image of our database. Each color on this representation corresponds to a single orientation from 0 to 360°. On the right of the same figure 4, one can see a representation of the curvature with the scale used to create this image. The red part of the scale is for high curves and the blue one is for the low curves [2].

IV. Log-polar curvelets based Signature of handwritten samples

The Curvelets have been interpreted in terms of curvatures and orientations and have led to the construction of a compact representation (a compact signature) for each writing. For each point of handwritten shapes, we have a couple of value (Orientation, Curvature) which is projected in the 2D signature space.

Each pixel of the signature correspond to the number of occurrences of the couple (x,y), where *x* is a specific orientation and *y* a specific curvature, found during the evaluation of the orientations and the curvatures of the image, see figure 5.

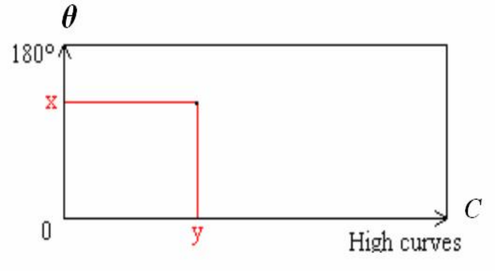


Fig. 5. Log-polar mapping for the representation of the couples (*C*,θ).

Figure 6 represents a result on a medieval sample of an historical manuscript (*Bastarda* Middle-Ages handwriting on 6.1) and a natural scene (on 6.2). One can see on the natural image that the signature is much noisier than the signature of the handwritten image. Orientations and curvature are not correlated as for the text.

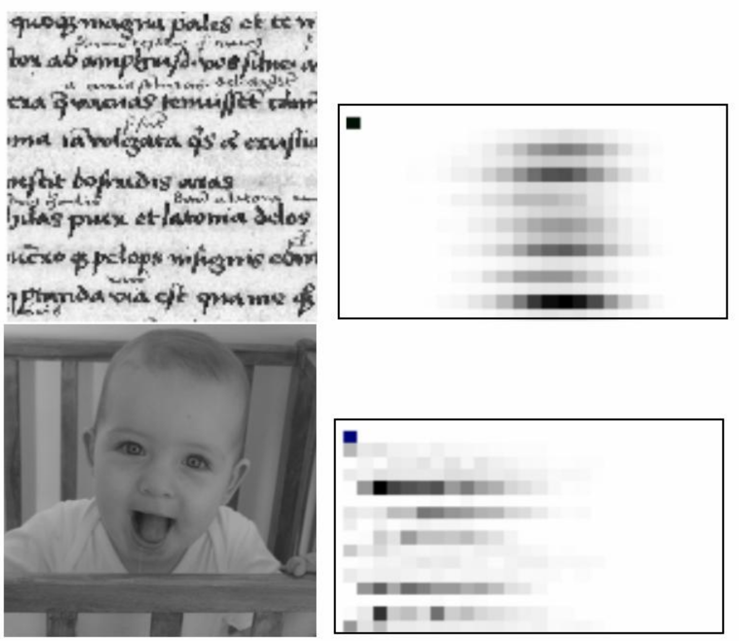


Fig. 6: A handwritten image with its signature (6.1) A natural image and its signature (6.2)

Instead of trying to compare two handwriting samples, we propose to compare their signatures with a new metric based on a *similarity* measure. The *similarity* measure is defined by the estimation of a *correlation* coefficient which indicates the similarity rates between each analysed samples. The expression of the correlation is:

$$r = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

where COV (X,Y) is the covariance between X and Y and σ is the standard deviation.

This measure is used to retrieve images from the database that have a similar writing style to the original request. Results are presented in a list of images having a decreasing correlation value with the request image, see figure 7. Under each image there are three different values. The first one is the correlation coefficient between the signature of the image and the signature of the requested one. The second value is the same correlation coefficient but with skeleton images. The last one is a linear combination between the two firsts values. It is on this last value that the retrieval is done.

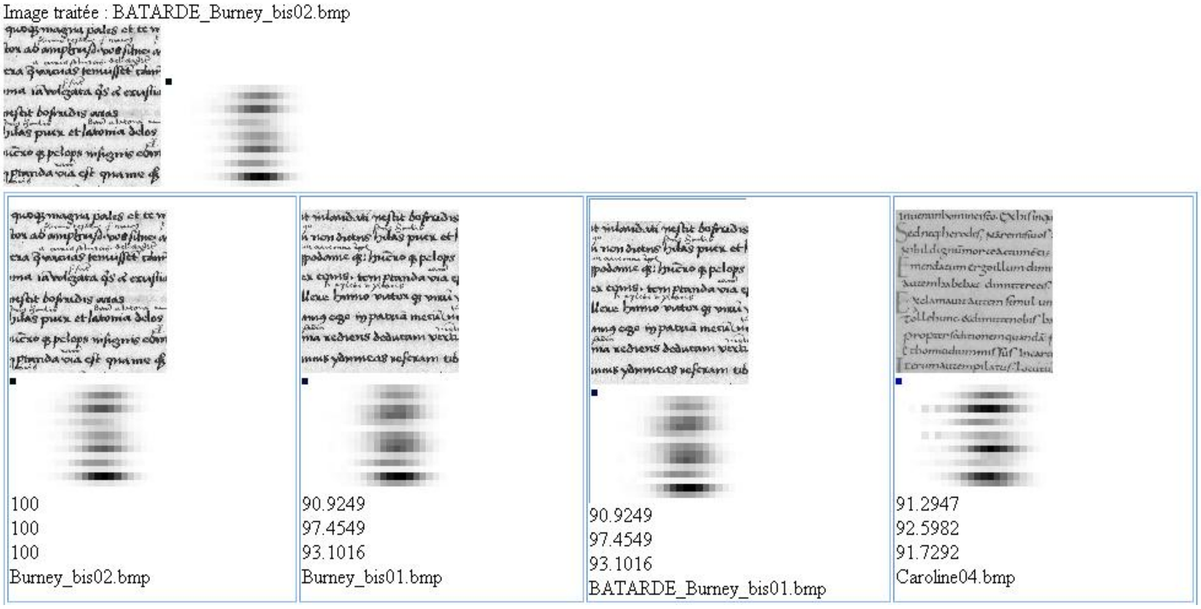


Image traitée : BATARDE_Burney_bis02.bmp

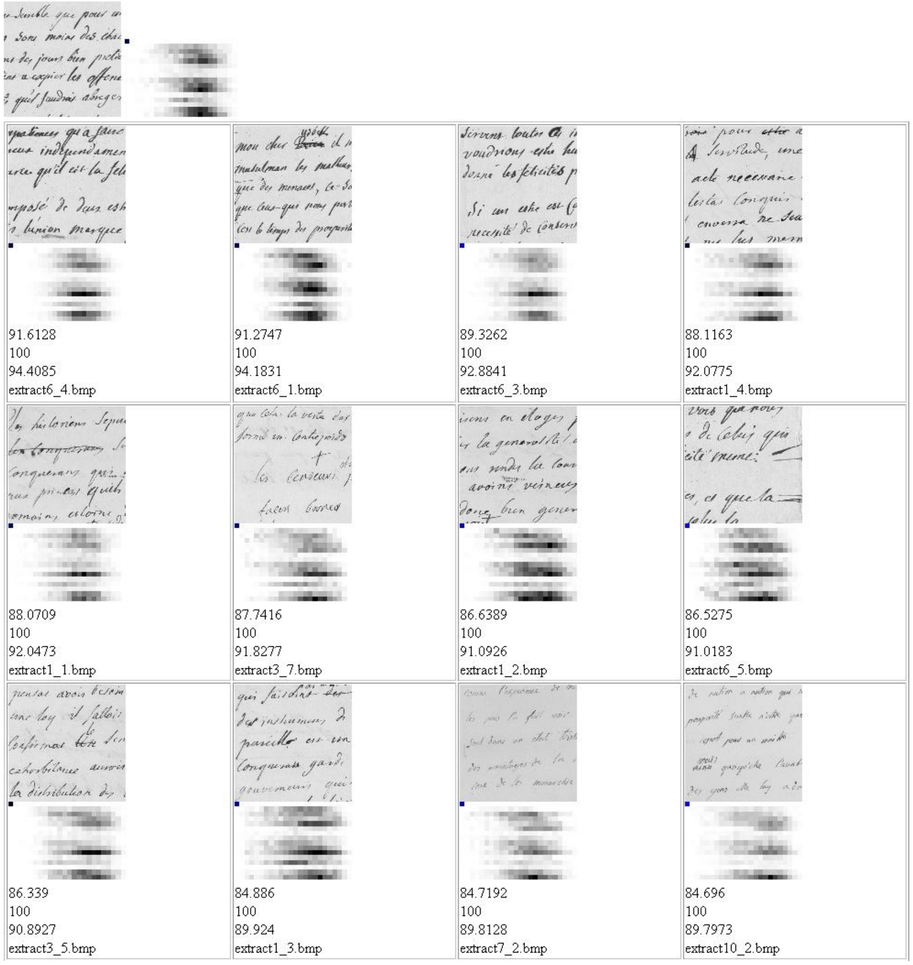
100	90.9249	90.9249	91.2947
100	97.4549	97.4549	92.5982
100	93.1016	93.1016	91.7292
Burney_bis02.bmp	Burney_bis01.bmp	BATARDE_Burney_bis01.bmp	Caroline04.bmp

Fig. 7. List of the four most similar retrieved images from the Middle Age database from a request *"Bastarda" handwriting sample.*

V. Experimental results and discussion

The tests are realized on two separated databases (Middle-Ages and Humanistic databases). The main experimental results are not very meaningful due to the size of our database (500 images per historical periods).

An example of responses of the system to a request from Montesquieu's database is illustrated in figure 8.



91.6128
100
94.4085
extract6_4.bmp

91.2747
100
94.1831
extract6_1.bmp

89.3262
100
92.8841
extract6_3.bmp

88.1163
100
92.0775
extract1_4.bmp

88.0709
100
92.0473
extract1_1.bmp

87.7416
100
91.8277
extract3_7.bmp

86.6389
100
91.0926
extract1_2.bmp

86.5275
100
91.0183
extract6_5.bmp

86.339
100
90.8927
extract3_5.bmp

84.886
100
89.924
extract1_3.bmp

84.7192
100
89.8128
extract7_2.bmp

84.696
100
89.7973
extract10_2.bmp

Fig. 8. List of the twelve most similar retrieved images from the Montesquieu database with an initial autograph request (from Montesquieu's hand).

The table 1 resumes the precision rates for different image requests of the combined Medieval and Humanistic databases. The precision is defined as the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. Results correspond to precision rate for a Top5, Top10 and Top 20.

	TOP 5	TOP10	TOP 20
Medieval and humanistic samples	94, 81 %	91,48 %	76,67 %

Table. 1. Precision of retrieval for the 5, 10 and 20 first answers for both humanistic and medieval database.

The precision decreases with the number of responses. Currently, we have in average 78 % of good retrieval (as precision) on the whole Middle-Ages database and 89% on the whole humanistic database.

Currently, the results on the Middle-Ages database are inferior: it is due to the difficulty to separate medieval handwriting style. We can notice that even the Paleographers have difficulties to find realistic separation between writing styles in close period of the history. For example, the differences between *Rotunda* (11th century) and *Bastarda* (11th century) is very difficult to establish. Our database is currently growing (with some digitization projects we are implied in) and we hope to get results still better.

This study is the result of a multidisciplinary collaboration: it has been led by expert advises and it has been designed with a specific image based analysis. We found many hopes in the continuation of such collaborations. But there is no doubt that an approach based exclusively on criterions of orientation and curvature extracted from the Curvelets transform would not be enough to solve classification and characterization problems on handwriting. In spite of that, we propose a track of very promising researches where the conjunction of two complementary criteria leads to a real discrimination between writings.

VI. Conclusion

This work has been done with the objective to help experts in paleography and literary experts: so it has been led by the idea of generics. On the one hand, the literary experts can find in this work a new tool to help them in their work of documents authentication and writer recognition. On the other hand, paleography experts can find a new tool to help them in their work of Middle-Ages handwritings classification. In theory, this method could also been used in any case where images are made of right segments and curved shapes like technical drawing, free-hand drawing, map, ... and more generally writing of all type.

Hence we can say with confidence that Curvelets Transforms can be used as a general technique for feature detection and is equivalent or in comparison to other algorithms described in literature. Furthermore we have discovered that Curvelets transforms in some cases give better results than Gabor transforms. This comparison is indirectly based on the papers on Gabor [12] for writer recognition. To prove that Curvelets features are really detecting the structure of writing we intent to used the two curved and oriented reconstructed versions of the handwriting (see figure 4) so as to find local similarities and writers invariants. This part is currently under study. Finally, we can extend the idea that we can join more than one type features into a single coherent feature set, which can be used for writer identification.

Acknowledgments

We want to thank the palaeographer Denis Muzerelle from the IRHT of Paris for his advices and recommendations for medieval handwritings classification and also Catherine Volpilhac-Auger from the CERPI of the ENS-Lettres of Lyon to have helped us to understand the genesis of the Montesquieu's corpus and to have guided to us towards the definition of robust criteria for styles differentiation.

References

[1] F. Aiolli, M. Simi, D. Sona, A. Sperduti, A. Starita, G. Zaccagnini, *"SPI: a System for Palaeographic Inspections. AIIA Notizie"*, http://www.dsi.unifi.it/AIIA/, p.34-38, Vol. 4, 1999.

[2] J.P. Antoine, L. Jacques, *"Measuring a cruvature radius with directional wavelets"*, Inst Phys Conf Series, p. 899-904, 2003.

[3] A. Bensefia, L. Heutte, T. Paquet, A. Nosary, *"Identification du scripteur par représentation graphèmes"*, CIFED'02, p.285-294, 2002.

[4] M. Bulacu, L. Schomaker, *"Writer style from oriented edge fragments"*, Computer Analysis of Images and Patterns (CAIP), p. 460-469, 2003.

[5] E. Candès, D. Donoho, *"Curvelets: A Surprisingly Effective Nonadaptive Representation of Objects with Edges"*,2000.

[6] S.H. Cha, S. Srihari, *"Multiple Feature Integration for Writer Verification"*, the 7th International workshop on Frontiers in Handwriting Recognition, IWFHR VII, p. 333-342, 2000.

[7] J.-P. Crettez, *"A set of handwriting families: style recognition "*, ICDAR 95, p.489-494, 1995.

[8] W. Kuckuck, *"Writer recognition by spectra analysis*", Int. Conf. In Security through Science Engineering, p. 1-3, 1980.

[9] U.V. Marti, R. Messerli, H. Bunke, *"Writer identification using text line based features"*, ICDAR'01, p.101-105, 2001.

[10] I. Moalla, F. LeBourgeois, H. Emptoz, A. M. Alimi, *"Contribution to the Discrimination of the Medieval Manuscript Texts: Application in the Palaeography"*, Lecture Notes in Computer Science, Vol. 3872, p.25-37, 2006.

[11] H.E.S Said, G.S. Peake, T.N. Tan, K.D. Baker, *"Writer identification from non-uniformly skewed handwriting Images"*, British Machine Vision Conference, p. 478-489, 1998.

[12] Shen, C., Ruan, X.G. and Mao, T.L., Writer identification using Gabor, 2002, Vol. 3, 2061-2064. [8, 13, 24]

[13] J.L STARCK, F. MURTAGH, E.J. CANDES, Gray and color Image contrast enhancement by curvelet transform, in IEEE Transactions on Image processing, pp. 706-717, vol. 12, No 6, 2003.