Curvelets based queries for CBIR Application in handwriting collections

G. Joutel, V. Eglin, S. Bres, H. Emptoz
LIRIS, *INSA-Lyon, F-69621, France*
{guillaume.joutel,veronique.eglin,stephane.bres,hubert.emptoz}@insa-lyon.fr

Abstract

This paper presents a new use of the Curvelet transform as a multiscale method for indexing linear singularities and curved handwritten shapes in documents images. As it belongs to the Wavelet family, this representation can be useful at several scales of details. The proposed scheme for handwritten shape characterization targets to detect oriented and curved fragments at different scales so as to compose an unique signature for each handwritten analyzed samples. In this way, Curvelets coefficients are used as a representation tool for handwriting when searching in large manuscripts databases by finding similar handwritten samples. Current results of ancient manuscripts retrieval are very promising with very satisfying precisions and recalls.

1. Introduction

This paper describes the construction of an automatic retrieval system for handwritten historical manuscripts as part of a regional project concerning the digitization of ancient handwritten collections. The system does not involve recognition and is used to query a dataset composed by two collections: 200 pages images from the Middle-Ages Latin manuscripts set and 200 pages images from the authors and humanistic manuscripts set.
The information is usually in numerical bitmap format and the collections are regularly increased by new digitized manuscripts images. In this study we only use samples of those collections. This amount of data makes it difficult to access or retrieve information. A basic approach consists in using textual metadata that are created manually in an expensive and tedious process. In Content Based Image Retrieval (CBIR) system, earlier approaches tend to require queries in the form of writing samples. Then the query is compared with other samples of the collection using a matching function. The retrieval is performed by matching in feature space derived from images in the test collection. In this work, we are interested in digitized Middle-Age documents (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). This work is dedicated to palaeographers and historians. It intends to help them in their work of manuscripts dating, expertise and authentication. In that context, we need an original approach to answer the opposite questions raised by those two kinds of documents from the *Middle-Ages* and the *humanistic* collections. For the images retrieval of Middle-Ages manuscripts, the objective is to find writer *independent* and style *dependent* primitives. For a special class of handwritings, the *Gothic* family for example, differences between copyists' writings are too small to give any relevant information for the dating process. In opposite, for humanistic manuscripts, the objective is to find writer *dependent* primitives for a writer identification task. Considering the fact that palaeographers judge that *curvature* and *orientation* are two fundamental characteristics of handwritings, we have searched a way to compute them on the writing shapes for both collections. To do that, we have developed a methodology:

- Sensitive to variations at the frontiers of shapes,
- Sensitive to the evolution of these criteria at different scales,
- Revealing the variability of shapes and their anisotropy,
- Robust to disturbed environments (presence of disturbed backgrounds, of partial shapes…),
- That doesn't require prohibitive costs of treatment nor great storage volume for each analyzed page image.

Our choice relates to a redundant multi-scale transform: the Curvelet transform that is more robust than wavelets for the representation of shapes anisotropy, of *lines segments* and *curves* in the images. Standard wavelet-based tools suffer of their incapacity to locate thin structure of lines and thin variations all along curves. The paper is organized in three parts: a short presentation of related works in CBIR applications, then the presentation of the Curvelet transform and the construction of the individual signature for each handwriting image of the database. Finally, we present our experiments and discussed results that are displayed in precision and recall curves.

2. Overview of the system

Given a large collection of documents, one is always confronted with the problem of locating the desired information. Thus the task of image retrieval can be loosely described as effectively finding the documents which contain the information a user needs. This usually involves converting all documents and user's information need (the user is directly involved in exploring data set by specifying a query) into some internal representation (for indexing documents and queries) and then matching the documents and the queries over the representations. The user must then interpret results of the matching process.

Our system of retrieval by content depends on the notion of *Similarit*y based on the composition of an individual signature for each handwriting sample of the database, see figure 1 for the system overview.

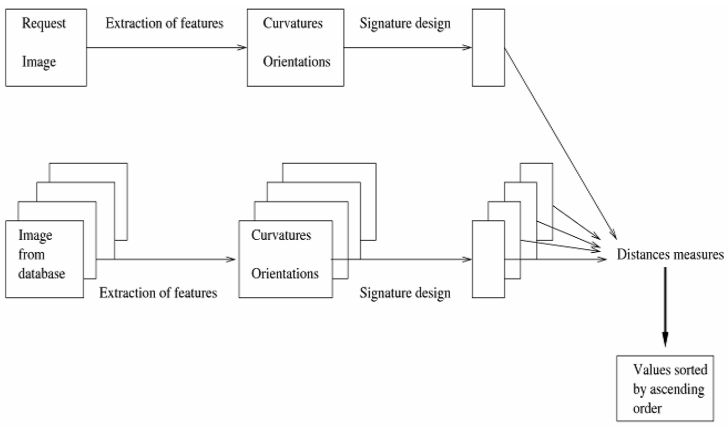


Fig. 1. Steps of our retrieval system.

The signature retrieval task involves two steps to process all signatures from the documents and then to perform comparisons on these signatures. Signature matching is based on a similarity measure defined as a normalized correlation. On considering then the first tops 5, 10 and 20 ranks, harmonic mean of 81.65 % on precision and recall shows a very promising accuracy of the retrieval technique and the feasibility of the Curvelets based tool to characterize various kinds of ancient handwriting samples. Our solution consists in extracting from the database all other documents from the same authors with the best possible recall.

3. Related works

The need for searching scanned handwritten documents are involved in application such as collections dating, works genesis, erudite study, critical edition, documents authentication… Recently Srihari et al. in [10-11] have realized the importance of handwritten document retrieval and have presented their retrieval system dedicated to forensics applications such as writer identification. Different distances are currently used to access the best matching between different set of handwriting samples, [6]. 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.

In handwriting classification which is our main goal, the most closely related works are [7] and [12]. In [7], authors propose an analysis of the variability of handwritings based on two observations: firstly the thickness of the tracing and the spatial density of characters and secondly the successive directions of the tracing. This work has been led on contemporary documents only. The only work that is relative to the analysis of the Middle-ages documents referenced in [1] proposes classifications procedures which are today debatable by palaeographers. Some recent studies have tried to develop more robust approaches for a non supervised classification. In [8] for example, authors are interested in ancient Latin and Arabic manuscripts of the Middle-Ages before 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 the letters frequencies.

It is difficult to pretend to be exhaustive in the description of handwritten shapes for the retrieval, 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, [9]. Within this kind of generic approach, it is possible to classify handwriting samples into visual distinct style classes.

As for now, studies for writing characterization are mainly 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 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. But features, which capture the global characteristics of the writer's individual writing styles, 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 allows combining both of them for the writers' style classification and the individualities characterization.

In our laboratory, we focuse on Frequency domain and signal decomposition (with Hermite and Gabor based features) to reveal salient oriented graphemes on high frequencies contours at different scales and resolutions. The only methods that use wavelets decomposition for handwritings classification do not make any hypothesis on the measured information, [9]. The general case is to compute several measures on wavelets coefficients but without any real link with the real meaning. In our proposition, we lead a Curvelets based analysis for a designed decomposition in curvature and orientation which have a real sense for palaeographers. It is important to remind that those measures of orientation and curvature are obtained without any segmentation of the handwriting. Our contribution differs from the cited works because it is dedicated to both Middle-Ages and contemporary handwritings documents. Moreover it uses a wavelets based tool, the Curvelets, which has never been used in this context.

4. Our method

4.1. The Curvelet transform

The Curvelet transform is a special member of the emerging family of multi scale geometric transforms, [5] It has been developed in the last few years in an attempt to overcome inherent limitations of traditional multi scale representations such as Wavelets. These limitations arise from the well-known and frequently depicted fact that the two-dimensional wavelet transform of images exhibits large wavelet coefficients even at fine scales, all along the important edges in the image, so that in a map of the largest wavelet coefficients one sees the edges of the images repeated from scale to scale. It means that many wavelet coefficients are required to reconstruct properly the edges in an image. Conceptually, the Curvelet transform is a multi scale pyramid with many directions and positions at each length scale, and needle-shaped elements at fine scales. This pyramid is nonstandard because Curvelets have geometric features that set them apart from wavelets and the likes. Curvelets obey a parabolic scaling relation which says that at scale 2^{-j}, each element has an envelope which is aligned along a ridge of length $2^{-j/2}$ and width 2^{-j}. Mathematically, if one works in R^2, Curvelets can be seen as follows: first, we have to consider a radial window $W(r)$ and an angular window $V(t)$ where r and t are polar coordinates in the frequency domain. These are both smooth, nonnegative and real-valued, with W taking positive real arguments and supported on $r\in[1/2,2]$ and V taking real arguments and supported on $t\in[-1,1]$. These windows will always obey the admissibility conditions:

$$\sum_{j=-\infty}^{\infty} W^2(2^j r) = 1, \qquad r \in (3/4, 3/2);$$

$$\sum_{\ell=-\infty}^{\infty} V^2(t-\ell) = 1, \qquad t \in (-1/2, 1/2)$$

Then, for each $j \geq j_0$, a frequency window U_j is defined in the Fourier domain by

$$U_j(r,\theta) = 2^{-3j/4} W(2^{-j}r)\, V(\frac{2^{\lfloor j/2 \rfloor}\theta}{2\pi}).$$

where $\lfloor j/2 \rfloor$ is the integer part of j/2. The support of U_j is a polar wedge with the support of W and V, the radial and angular windows applied with scale-dependent window widths in each direction, see fig. 2.

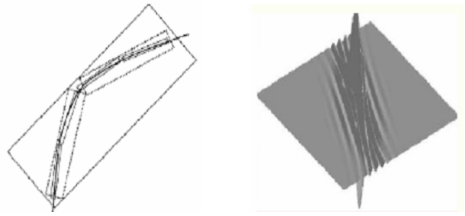


Fig. 2. Parabolic scaling law for a better anisotropy detection

All curvelets at scale 2^{-j} are then obtained by rotations and translations of the corresponding "mother" curvelets. Curvelets coefficients are then inner products between and element of the image and the curvelet basis. More details can be found in [5]. Once the Curvelet transform is computed for a handwriting image sample, we compute all curvature and orientation values that are present all along shapes contours.

4.2. Orientations and curvatures computation

The Curvelet transform gives us an analysis of pixels for different scales and orientations. One pixel on a curve can be potentially detected in several orientations depending of the curve. Each Curvelet coefficient corresponding to this pixel in a detected orientation is then compared to coefficients in all other detected orientations of the same pixel. We only retrieve the more significant coefficients values. Finally, the number of significant orientations gives us an evaluation of the overall curvature of a pixel, as it has been presented in [2] by J-P Antoine. Figure 3 shows an example of curvature computation on a sample of Montesquieu's handwritten shapes (1750). The colored scale (from blue to red) indicates the increasing values of curvature density.

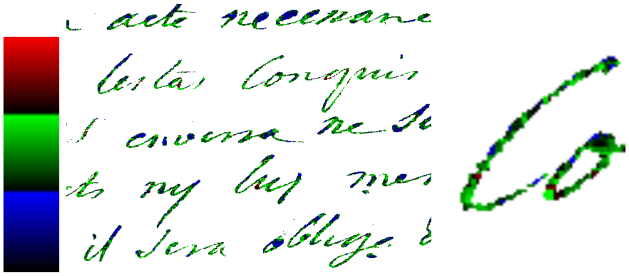


Fig. 3. Curvature representation from the Curvelets decomposition in a handwriting sample.

4.3. Signature construction

From the computation of orientations and curvatures, we define a signature for each handwriting manuscript. The signature is defined as the matrix of couple occurrences (curvature, orientation). Then, for each pixel of the manuscript image, we keep the couples of curvature and orientations values and we increment the corresponding elements with (curvature, orientations) coordinates in the matrix. Figure 4 presents an example of matrix plotted for a Middle-Ages handwriting image.

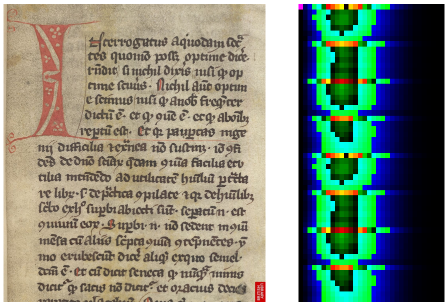


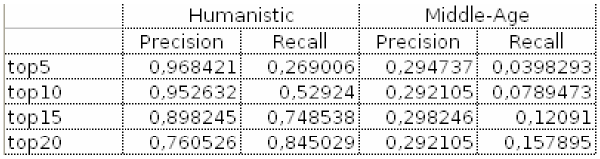
Fig. 4. Signature of a Middle-Ages handwriting sample.

In our retrieval system, the distance between the queried signature and each signature of the database is computed using a normalized correlation similarity. The similarity measure S(X,Y) between two images X and Y is defined as follows: S(X,Y)=Cov(X,Y)/σxσy, where *cov(X,Y)* is the covariance between *X* and *Y* and σx and σy the standard deviations.

5. Experiments and results

In this part, the test setup and the experimental results obtained for the image retrieval task are described. In the test setup for manuscript images retrieval, the images are divided into two groups: one group is composed by known handwriting images and the second group is composed by queried images for testing. We have tried to test our retrieval system on two databases (400 images): the humanistic and the Middle-Ages databases (composed by manuscripts of the British Library, http://www.bl.uk and the Digital Scriptorium, http://sunsite.berkeley.edu) In retrieval system, the performance is subjective and its evaluation is difficult, especially for the Middle-Ages base where we do not have precise information concerning Middle-Ages European handwriting dating process. The set of known images is selected and the image retrieval process is carried out against this set of known handwritings. In each case, we compute the average precision and the recall values for each database. Precision and recall values are obtained from the confusion matrix of relevant and irrelevant returned answers (as results images) for a given query (as initial image).

The ranks of the handwritings belonging to the correct authors (for the humanistic base) and to the correct styles (for the Middle-Ages base) of a queried handwriting are noted in Table 1.



	Humanistic		Middle-Age	
	Precision	Recall	Precision	Recall
top5	0,968421	0,269006	0,294737	0,0398293
top10	0,952632	0,52924	0,292105	0,0789473
top15	0,898245	0,748538	0,298246	0,12091
top20	0,760526	0,845029	0,292105	0,157895

Table 1. Precision and recall values for the first ranks for the two databases.

In this table, we have precisely shown results concerning French authors' handwritings (Flaubert, Montesquieu…) and 12th to 15th century European handwritings. For the humanistic database, results show that on average 4.84 images on 5 are handwriting samples made by the same writer as the queried handwriting. In the top 15, a precision of 89.82 % is obtained for a recall of 74.85%. This implies that, on average, 74.85% of all images of the correct author of a queried image were returned in the first 15 answers. Figure 5 presents the precision and recall graph and the first 10 answers of the system for a given query in the humanistic database.

We have also used a F-measure which combines the precision and the recall values (F=2P.R/(P+R)) and provides a single measure of the retrieval accuracy.

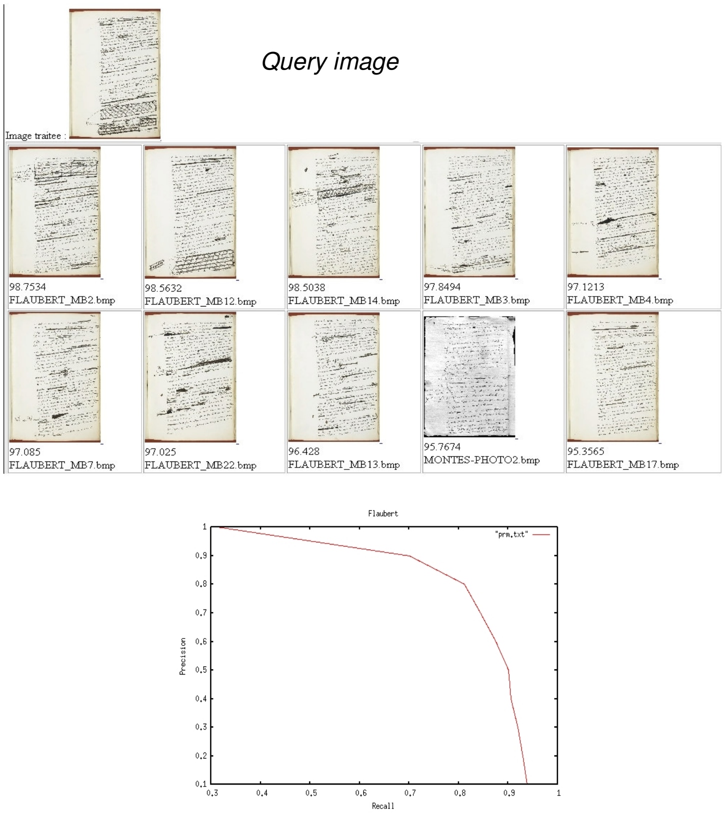


Fig.5. Example of answers of our retrieval system. Precision and recall graph for the humanistic database.

In the top 15 ranks, the F-measure is 81.65 (P=89.82 and R=74.85). For the Middle-Ages base this value is 43.41.

The second thing to test is the robustness of our system to classical noise degradations. It has been shown in [5] that for denoising Curvelets are well adapted but our wondering was about the robustness of our signature to this noise without any denoising. We have tested degradations of several kinds as: resolution changes, scale changes, zooms over part of the text, Gaussian noise, word reproducing in the page, dilatations and erosions. In the 33 images created by these degradations, the 25 firsts one had a retrieval rate over 90%.

6. Conclusion

This study is a part of software platform dedicated to paleographers and literary experts, to help them in their work of manuscripts dating and authentication through different historical periods. The methods that are embedded in the retrieval system have been thought with the idea of generics: in theory, they can be used in any case where images are made of right segments and curved shapes like technical and free-hand drawings, maps, and more generally writings of all types.

Hence, we can say with confidence that Curvelet transforms can be used as a general technique for feature detection. Furthermore, we have discovered with the promising recall values of our system that Curvelet transforms in some cases give better results than Gabor transforms. This comparison is indirectly based on the results on Gabor based writer recognition, [9]. To prove that Curvelets features are really able to detect the structure of written shapes we intent to use both 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.

Acknowledgments: We want to thank the palaeographer Denis Muzerelle from the IRHT of Paris for his advices for medieval handwritings classification and also Catherine Volpilhac-Auger from the CERPI of the ENS-Lettres of Lyon for her definition of robust criteria for styles differentiation.

7. References

[1] F. Aiolli, M. Simi, D. Sona, A. Sperduti, A. Starita, G. Zaccagnini, *"SPI: a System for Palaeographic Inspections. AIIA Notizie"*, 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"*,in Schumaker L., *Curves and surfaces filtering,*Vanderbilt University Press, 1999.

[6] S.H. Cha, S. Srihari, *"Multiple Feature Integration for Writer Verification"*, IWFHR VII, p. 333-342, 2000.

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

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

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

[10] Zhang, B., Srihari, S.N., Binary vector dissimilarity measures for handwriting identification, in Document recognition and Retrieval, SPIE, 5010 pp.28-38, 2003.

[11] Zhang, B., Srihari, S.N., Word image retrieval using binary features, in Document recognition and Retrieval, SPIE, 5296, pp.45-53, 2003.