Contour Based Features for the Classification of Ancient Manuscripts

Imran SIDDIQI, Florence CLOPPET, Nicole VINCENT
Laboratoire CRIP5 – SIP, Paris Descartes University
75006, Paris, FRANCE
{siddiqi, florence.cloppet, nicole.vincent}@math-info.univ-paris5.fr

Abstract. This paper presents an effective system for the classification of ancient handwritten documents according to the writing style. We have employed a set of features that are extracted from the contours of the handwritten images. These features are based on the direction and curvature histograms that are extracted at a global level from local contour observations. Two writings are then compared by computing the distance between their respective histograms. The system evaluated on medieval texts exhibits promising results.

1. Introduction

Writing is an inseparable component of any culture and the evolution of the style and form of writing over time reflects the historical and cultural changes of society. Knowledge about individual letter shapes, ligatures, punctuations and abbreviations and, the way they have evolved, enables the palaeographers and historians identify the periods and the geographical place in which a manuscript was written. The quantity of these ancient manuscripts stored in archives, libraries and private collections is enormous and it will be useful to develop a computer system that could help the palaeographers in manuscripts dating, classification and authentication. In the present study, we will be particularly interested in the digitized medieval handwritten texts. We will focus our interest on finding a set of features from an image which could be helpful in automatically classifying the huge bases of manuscripts. The analysis and classification writing styles, which is the main objective of this work, is similar in many respects to the recognition of writers, the latter requiring more precision in the decision to assign a script to a particular class.

Among the well known methods for handwriting classification (Crettez, 1995) proposes an analysis of the variability of handwritings with the objective of identifying the family of the handwriting style. The fractal analysis of a handwritten image reflects the writing style of its author and serves to classify writings according to their legibility (Vincent & al., 2000). The effectiveness of redundant patterns of a writing for characterizing its author has been depicted in (Siddiqi & Vincent, 2007). These methods have been validated on contemporary writings. Important contributions on medieval and humanistic manuscripts include (Aiolli & al., 1999; Eglin & al., 2007; Joutel & al., 2008; Yosef & al., 2004; Moalla & al., 2006). Each of these studies focuses on one specific tool motivated by the characteristics of the writing (time, location, alphabet, language etc.). In our case, we are concerned by a wider variety of writings thus we need to define more generic features compared to the ones presented in the literature. We present a system for automatic classification of writings based on the orientation and curvature histograms that are computed from a handwritten sample. We mainly focus on a sample of 310 medieval manuscripts selected from the collection of IRHT[1]. To perform training and evaluation, we have split these documents into roughly two equal parts, the first half contributing to the training set while the other to the test set. The methodology is detailed in the sections to follow.

2. Proposed Method

We now present our method and its application to the classification of writings. We need to define a set of features that capture the writing style but are independent of the writing instrument. The instrument dependency could be eliminated by working either on the contours or on the skeleton of writing. We have chosen to work on contours as skeletonization eliminates the writer and writing style dependent variations between the character shapes and is more useful in handwriting recognition. On the contrary, for the classification of writing styles, these variations are important and need to be preserved. Thus we introduce a number of features that are based on the contour of the handwritten text images.

2.1 Feature Extraction

In order to extract the contours, the document images first need to be binarized. However, the quality of historical manuscripts is generally quite poor as the documents degrade over time due to, for instance, storage conditions. Thus the foreground and background are difficult to separate and the classical thresholding methods fail when applied to these documents (Leedham, 2003). We therefore chose to employ the binarization scheme

[1] This study has been carried out in the framework of the project ANR GRAPHEM: ANR-07-MDCO-006-04.

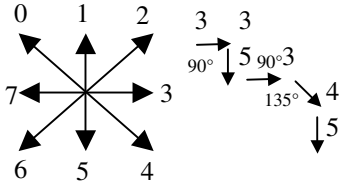
presented in (Yosef, 2005) which claims to work quite well on degraded images. First a global thresholding is carried out that enables separating the background from noise-free characters. A more sophisticated local method is then employed to binarize the noisy characters. Figure 1 shows an example of binarizing a noisy image, the algorithmic details could be found in (Yosef, 2005). Once binarized, we extract the connected components in the image (using 8-connectivity) and for each of the components we find its contours and the corresponding Freeman chain code.

(a) (b) (c)

Figure 1: a) Original Image b) Global Thresholding c) Final Binarized Image

It is known that humans can distinguish two writings instinctively and the most important element they are sensitive to, is the overall orientation of the writing. This orientation information can be captured by computing the slope density function (histogram of all the slopes/chain codes **f1**) of the contours in a handwritten text image. The bins of the histogram represent the percentage contributions of the principal stroke directions: horizontal, vertical, left-diagonal and right-diagonal. Since the images are offline, we do not have the drawing order of the strokes hence whether a stroke is considered forward or backward is dependent on the way the contour is traced.

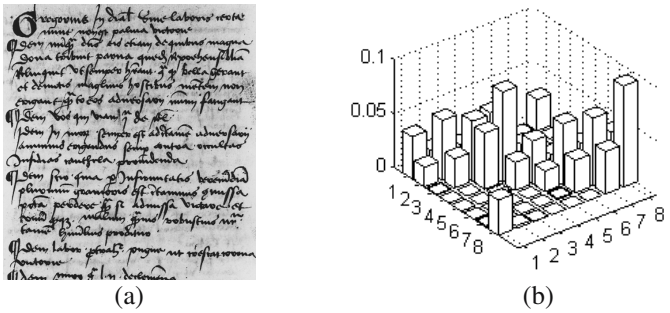
Another important visual aspect of handwriting is the curvature which can be estimated by the differential chain codes and that is linked to the way the muscles are conducted by the writer. The differential codes are computed by subtracting each element of the chain code from the previous one and taking the result modulo *d* (connectivity). Their histogram, also known as the curvature density function (**f2**), could be used to characterize the handwritten sample. Figure 2 shows the eight principal directions and the slopes and curvatures associated with the chain codes and their differentials.



Slope	*Chain Code* c_i	0	1	2	3	4	5	6	7
	θ_i	135°	90°	45°	0°	315°	270°	225°	180°
Curvature	$(c_{i+1}-c_i)$ *mod 8*	0	1	2	3	-	5	6	7
	$\Delta\theta_i$	180°	135°	90°	45°	-	315°	270°	225°

Figure 2: Principal directions and the angles associated with chain codes and their differentials

The slope and curvature histograms could serve well to perform a crude classification but are insufficient to capture the fine details in writing. We thus propose to count not only the occurrences of the individual chain code directions but also the chain code pairs. We scan the chain code sequence and for each pair *(i,j)* we increment the bin *(i,j)* of the histogram **f3**. We next employ the same principle on chain code triplets, that is; we define a three dimensional histogram (**f4**) where the bin *(i,j,k)* of the histogram represents the percentage contribution of the triplet *i,j,k* in the chain code sequence of the contour of a handwritten image. Of course the two matrices (8x8 and 8x8x8) are quite sparse as all the possible combinations (of pairs and triplets) do not exist. Certain combinations are not possible while some are less probable. However, for the simplicity of calculations, we chose to keep the complete matrices. Figure 3 illustrates an example image and the corresponding histogram of code pairs.



(a) (b)

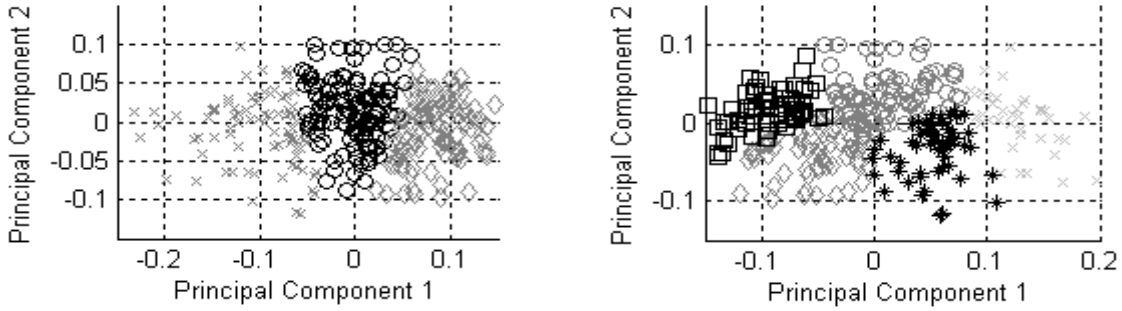
Figure 3: a) Input Image b) Corresponding histogram of chain code pairs

A set of four features (normalized histograms) is thus extracted from a handwritten sample as summarized in Table 1 along with the dimensionality of each. We next need to group similar writings into clusters, For this

we employ a k-means clustering algorithm where two writings are compared using a (dis)similarity measure defined on their respective features. We tested a number of distance measures including: Euclidean, $\chi 2$, Bhattacharyya and Hamming distance, $\chi 2$ distance reading the best results in our evaluations. Since the palaeographers presently search for a "*good*" value of *k* (number of classes), we have made *k* vary from a minimum of 2 to 25. Applying a PCA and reducing the writing representation to two dimensions, the writing classes obtained have been visualized in figure 4.

Table 1 Proposed features and their dimensionality

Feature	Description	Dim
f1	Histogram of chain code: *Slope Density function*	8
f2	Histogram of differential chain code: *Curvature Density function*	7
f3	Histogram of chain code pairs	64
f4	Histogram of chain code triplets	512



(a) (b)

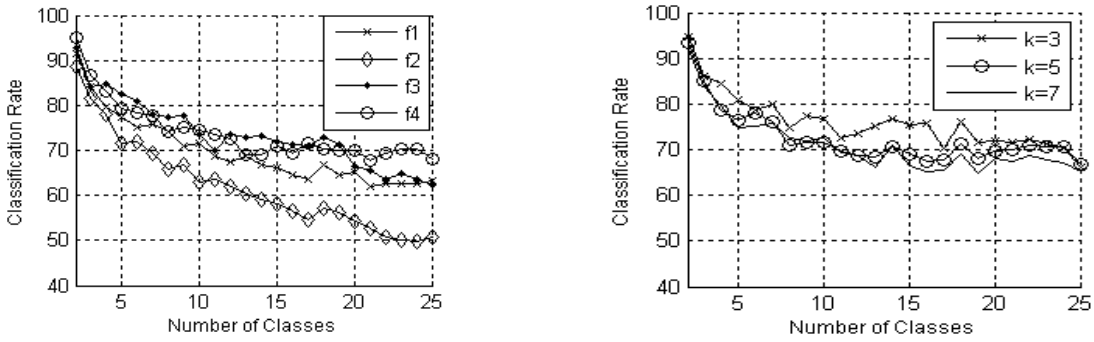
Figure 4: Writing Classes obtained with k-means for a) k=3 b) k=5

2.2. Classification

The objective of the classification is to pick up each image in the test set and assign it to one of the *k* potentially overlapping classes. An image is said to be correctly classified if it is attributed to the same class as that of its counter part in the training set, i.e. the two halves (training & test) of an image should belong to the same class to be considered as correct classification. The classification is carried out using the k-nearest neighbours (knn) with k=3,5,7 on the individual features as well as on their combination. The combination is performed by computing the distance between two writings as a weighted average of the distances between the individual features, the weights being assigned relative to the performance of individual features. It should also be noted that a k means with k=1 (i.e.; no clustering of the training set) and then a knn with k=1 changes the problem of classification to writer identification which will also be addressed in the experiments.

3. Experimental Results

We first present the classification results on individual features varying the number of classes. As it can be noticed from figure 5, for a smaller number of classes, there is not much difference between the performances of the four features with **f4** outperforming the rests. As the number of classes increases, the performance of **f4** becomes more or less stable while that of **f2** drops the most significantly. Combining these features naturally serve to enhance the classification rates as illustrated in figure 5b where the classification curves have been shown for three different values of k (in knn).



(a) (b)

Figure 5: Classification performance (KNN) a) On individual features with k=3 b) On combined features (f1-f4) with k=3,5 and 7

It should be noticed that the clusters of writings that we obtain by employing the proposed features have yet to be compared with the ones suggested by the palaeographers. The evaluation procedure that we followed (dividing the images into two disjoint sets) was meant to present the results in a quantified form. Since the results are very promising, we expect these features to be useful for the palaeographers.

We also tested the proposed features to perform writer identification and verification on the same data set. Writer Identification is performed by computing the distance between a query image from the test data and all the images in the training data, the writer being identified as the writer of the document that reports the minimum distance. For writer verification, we compute the Receiver Operating Characteristic (ROC) curves by varying the acceptance threshold, verification performance being quantified by the Equal Error Rate (EER): the point on the curve where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). Table 2 summarizes the system performance on writer recognition (numbers represent percentages), while the corresponding ROC curves have been illustrated in figure 6. We realize an identification rate of as high as 94% and an equal error rate of as low as 2.27% (with f4).

Table 2 Writer Recognition Performance

Feature	Top1	Top10	EER
f1	38	81	8.18
f2	35	79	9.52
f3	86	97	3.60
f4	94	99	2.27
f1-f2	68	89	5.87
f1-f4	93	99	2.59

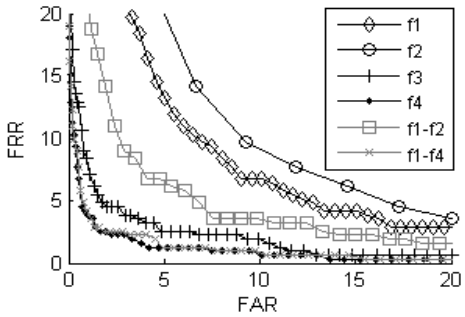


Figure 6: ROC Curves for writer verification

4. Conclusion

We have presented an effective method for the automatic classification of medieval manuscripts. The method is based on a set of features extracted from the contours of the handwritten text images. The classification rates achieved are very encouraging and validate the methodology put forward in this paper. In addition, the results of the system on writer identification and verification show its effectiveness for forensic applications as well. The proposed method is quite generic and can be applied to contemporary writings as well as the non-Latin languages such as Asian or Arabic scripts. Our future study will focus on comparing our classification with the one proposed by the palaeographers.

Acknowledgments

We would like to thank our colleagues from the IRHT (Institut de Recherche en Histoire des Textes) and the Ecole des Chartes for providing us with the images and exchanging fructuous views.

References

Aiolli, F., Simi, M., Sona, D., Sperduti, A., Starita, A., Zaccagnini, G. (1999). SPI: a System for Palaeographic Inspections. AIIA Notizie 4, 34–38.

Crettez, J.-P. (1995). A set of handwriting families: style recognition. In: Int'l Connference on Document Analysis and Recognition, pp. 489–494.

Eglin V., Bres S.,Rivero-Moreno C.J. (2007). Hermite and Gabor transforms for noise reduction and handwriting classification in ancient manuscripts. IJDAR 9(2-4):101-122.

Joutel, J.G., Eglin, V., Emptoz, H.(2008). A complete pyramidal geometrical scheme for text based image description and retrieval. International Conference on Image and Signal Processing 2008, Springer Verlag ed. Cherbourg-Octeville, Normandy, France.

Leedham, G., Yan, C., Takru, K., Tan, J.H.N., Mian, L.(2003). Comparison of Some Thresholding Algorithms for Text/Background Segmentation in Difficult Document Images. In Proc. Of ICDAR 2003. pp. 859-864.

Moalla, I., LeBourgeois, F., Emptoz, H., Alimi, A.M.(2006). Contribution to the Discrimination of the Medieval Manuscript Texts. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS,vol. 3872, pp. 25–37. Springer, Heidelberg.

Siddiqi, I., Vincent, N.(2007). Writer Identification in Handwritten Documents, In Proc. of ICDAR 2007, Curitiba, Brazil, pp. 108-112.

Vincent N., Boulétreau V., Sabourin R., Emptoz H. (2000). How to use fractal dimensions to qualify writings and writers, Fractals, World Scientific, Vol 8, n°1, pp.85-97.

Yosef I. B., Kedem K., Dinstein I., Beit-Arie M., Engel E. (2004). Classification of Hebrew Calligraphic Handwriting Styles: Preliminary Results, DIAL, pp 299-305.

Yosef I. B. (2005). Input sensitive thresholding for ancient Hebrew manuscript. Pattern Recognition Letters 26(8), pp 1168-1173.