2009 10th International Conference on Document Analysis and Recognition

A Set of Chain Code Based Features for Writer Recognition

Imran Siddiqi, Nicole Vincent
Paris Descartes University, Laboratoire CRIP5 – SIP
45, rue des Saints-Pères, 75006 Paris, France
{siddiqi, nicole.vincent}@math-info.univ-paris5.fr

Abstract

This communication presents an effective method for writer recognition in handwritten documents. We have introduced a set of features that are extracted from the contours of handwritten images at different observation levels. At the global level, we extract the histograms of the chain code, the first and second order differential chain codes and, the histogram of the curvature indices at each point of the contour of handwriting. At the local level, the handwritten text is divided into a large number of small adaptive windows and within each window the contribution of each of the eight directions (and their differentials) is counted in the corresponding histograms. Two writings are then compared by computing the distances between their respective histograms. The system trained and tested on two different data sets of 650 and 225 writers respectively, exhibited promising results on writer identification and verification.

1. Introduction

The need to recognize the writer of a handwritten document is a recurrent problem not only from the perspective of behavioral biometrics [2,10,14] but also in the context of handwriting recognition [9] exploiting the principle of adaptation of the system to the type of writer. Writer recognition is generally distinguished into writer identification and verification. Writer identification involves a one-to-many search where, given a document of an unknown authorship, the objective is to find its author in a reference base with documents of known writers. Writer verification on the other hand is a one-to-one comparison where, given two handwriting samples, one would like to determine whether the two samples have been written by the same person or not.

The early research in writer recognition has mainly witnessed the text-dependent methods where the two writing samples to be compared require to contain the same fixed text for example; signature verification. A few relatively recent studies [14,15] also present text-dependent writer identification systems. The text-independent methods on the other hand identify the writer of a document independent of its semantic content thus they are less constrained and more useful for practical applications. Another traditional classification of writer recognition methods is into global and local approaches. The global methods [3,11] are based on the overall look and feel of the writing whereas the local techniques [2,4] identify the writer based on localized features, which are inherent in the way a writer specifically writes characters. The latest trend in writer recognition is to use a set of patterns to which the actual writing is compared [2,13]. Combining the global and local features is also known to improve the writer recognition performance [5,14] and our research is inspired by the same idea.

We present a system for offline writer recognition using very simple features as recognition of the author can be done by human very instinctively. Human is mostly sensitive to contours and changes so we work on the contours of handwritten text images. We start with a global analysis of handwriting using the classically known histograms of chain code and their differential forms. We then propose their local variants that are calculated from small segments of handwritten text. Finally we perform a comparative evaluation of the two and explore their various combinations. The method has been detailed in the sections to follow.

2. Feature Extraction

In this section we present the proposed features and their extraction methods. Based on the hypothesis that the contour of a handwritten sample encapsulates the writing style of its author, we introduce a number of features that are based on the contour of the handwritten text images. We have chosen contour

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.136

IEEE computer society

instead of skeleton as skeletionization introduces some loss of information and is more useful in handwriting recognition where the writer-dependent variations between the character shapes need to be eliminated. Writer recognition on the other hand relies on these variations which are preserved by the contours. Starting with the initial gray-scale images of handwritten documents, we binarize them using Otsu's global thresholding algorithm and perform connected component extraction (using 8-connectivity). For each of these components, we then find its contours, each contour ($Contour_i$) being a sequence of consecutive boundary points:

$$Contour_i = \left\{ p_j \middle| j \le M_{i,} p_1 = p_{M_i} \right\}$$

With M_i being the length of contour i. We then calculate the Freeman chain code associated with each contour, the sequence $\{c_j | \ j \le M_i\text{-}1\}$ where $c_j \in \{0,1,\ldots,7\}$. The boundary pixels in the original binary image I are then labeled by their respective codes. We then proceed to the extraction of features from the newly formed image I^c. The contours are analyzed both at the global and local levels. At global level, to remove errors due to a false ordering of the pixels we employ the histograms of chain codes and their variants. At local level, we analyze small handwritten fragments. Finally the set of extracted histograms is used to characterize a handwritten sample.

2.1 Global Features

Chain code histograms have shown effective performance for shape registration and object recognition and since the handwritten characters issued by a particular writer have a specific shape, the histogram of the chain code calculated on the character contours is likely to capture directional information of its writer. The (8-bin) histogram of chain code ***f1*** is computed from image I^c representing the principal stroke directions: horizontal, vertical, left-diagonal and right-diagonal. Each bin of the histogram thus represents the percentage of the respective direction in an individual's writing. Since the images are offline, forward and backward strokes cannot be distinguished and are linked to the way a contour is traced.

The histogram is invariant towards different deformations but the most obvious limitation of the chain code histogram is that two totally different shapes can have similar histograms. This problem is dealt with by encoding not only the relative direction, but also the differences in successive directions: differential chain codes, computed by subtracting each element of the chain code from the previous one and taking the result modulo d, where d is the connectivity. Thus we get more information on the contour curve. The differential chain code at pixel p_i represents the angle θ_i (indexed as in figure 1) between the vectors $p_{i-1}p_i$ and p_ip_{i+1} and their distribution ***f2*** is used as the second feature to represent a handwritten text. Employing the same principle, we also compute the histogram ***f3*** from the second order derivative of the chain code C.

$(c_{i+1}\text{-}c_i)$ mod 8	0	1	2	3	5	6	7
θ_i	180°	135°	90°	45°	315°	270°	225°

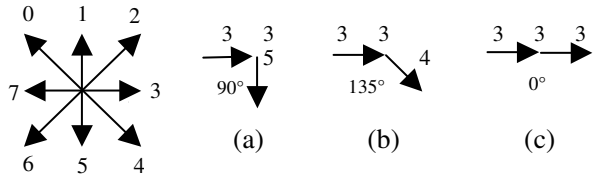

Figure 1. Differential chain codes and corresponding angles

To get some visual feeling from the text we can also study curvature, it is very much linked with the physiological way the strokes are written from the strength involved by the muscles and the way they are operated. Of course curvature can be deduced from the derivative but we approximate the curvature index at each point of the contour using the histogram based algorithm proposed in [1]. For each c_j in image I^c, we take K forward and K backward neighbors (K being linked to the height of the character base line and fixed to 7 in our case) and calculate two histograms ($f(j)$ and $b(j)$) representing the orientation of the segments on both sides of c_j (figure 2). The curvature index at c_j is then estimated by the reciprocal of the correlation coefficient between the forward and the backward histograms.

$$\rho_{inv} = \frac{\sum_{i=0}^{7} (f(i) - m_f)(b(i) - m_b)}{\sum_{i=0}^{7} (f(i) - m_f)^2 \sum_{i=0}^{7} (b(i) - m_b)^2}$$

With f and b being the forward and the backward histograms while m_f and m_b their mean values respectively. A high value (close to 1) of the correlation coefficient characterizes similar histograms and hence a low curvature index and vice versa. The correlation coefficients are calculated for each point of the contour sequence and are counted in a histogram ***f4***.

After having defined the four histograms that bring some global information on the directions of the

strokes, the evolution of the directions and on the curvature of the drawings, we now introduce some local features.

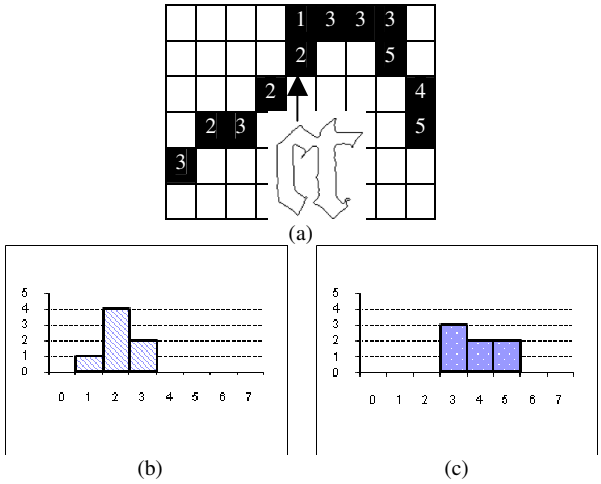

Figure 2. a) Contour pixel p_j (with code c_j) of a character b) Backward histogram at p_j c) Forward histogram at p_j

2.2 Local Features

At the local level, we aim to analyze small contour fragments. In different studies these fragments are chosen in such a way that they carry some semantic information but we think that writer recognition can be performed without the decipherment of the characters. So the fragments we consider are parts of the handwritten text image *I* contained in small windows. The image is first divided into a large number of small windows of size *nxn* employing the window positioning algorithm presented in [13] in order to position the windows in an adaptive way with respect to the writing. The window size *n* is chosen empirically on a validation set and is fixed to 13x13 in our case. The windows positioned over a text image and the corresponding contour image have been illustrated in figure 3.

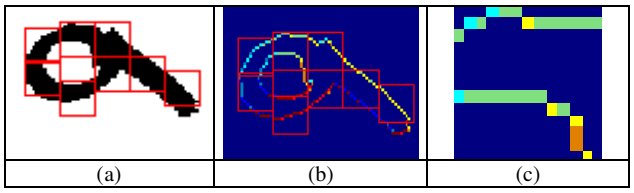

Figure 3. a) Windows positioned over text b) Windows positioned over the contour image c) Directional distribution in a window

For each window, we compute the distribution of the 8 directions with respect to the total segment length within the window, the percentages being clustered into *p* (set to 10:[0-10], [10-20],…, [90-100]) intervals. We then build an accumulator (stroke direction histogram) which is a two dimensional *dxp* array where *d* is the connectivity (8 directions). The accumulator is initialized with all bins set to zero. For each window *w*, containing the chain code sequence C^w, the bins *(i,j)* of the histogram (accumulator) are incremented by 1 if the direction *i* is represented in the j^{th} cluster, where *j* is given by:

$$j = \frac{card\,(C_i^w)}{card\,(C^w)\times p}\times 100 \quad With: C_i^W = \left\{c_j \in C^w \middle| c_j = i\right\}$$

Using the same principle, we calculate the histograms (matrices) ***f6*** and ***f7***, superimposing the windows on images $I^{c'}$ and $I^{c''}$ generated by labeling the contour pixels by their first and second order differential chain codes respectively.

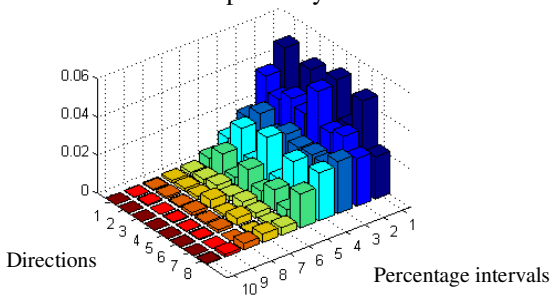

Figure 4. Stroke Direction Histogram of a handwritten image

We thus extract a set of seven features (histograms): four (f1 – f4) of them computed globally while three (f5 – f7) calculated locally. Table 1 summarizes the proposed features along with the dimensionality of each.

Table 1 Proposed features and their dimensionality

Feature	Description	Dim
f1	Chain code histogram	8
f2	1st order differential chain code histogram	7
f3	2nd order differential chain code histogram	8
f4	Curvature Index histogram	11
f5	Local stroke direction histogram	80
f6	*f2* computed locally	70
f7	*f3* computed locally	80

3. Writer Recognition

Once the handwriting samples have been represented by their respective features, we need to compute the distances between respective features to define a (dis)similarity between two handwriting samples. We tested a number of distance measures including:

Euclidean, χ^2, Bhattacharyya and Hamming distance, χ^2 distance reading the best results in our evaluations.

Writer Identification is performed by computing the distance between the query image Q and all the images in the data set, the writer of Q being identified as the writer of the document that reports the minimum distance (knn with k=1). For writer verification, the Receiver Operating Characteristic (ROC) curves are computed 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). The identification and verification results are presented in the following section.

4. Experimental Results

For the experimental study of our system, we have chosen two different data sets: IAM [7] and RIMES [6] which contain samples of unconstrained handwritten text. The IAM database comprises handwritten forms with English text of variable content while the database RIMES contains handwritten letters in French text. The texts in both data sets have been scanned at a resolution of 300 dpi and digitized to 256 gray levels. We have chosen a subset of 225 writers from the RIMES data set while for the IAM data set, we have kept the first two images for the writers having more than two pages and split the image roughly in half for writers who contributed a single page thus ensuring two images per writer, one used in training while the other in testing.

Table 2 Writer Recognition on Individual Features

Data Set / Feature	IAM 650 Writers			RIMES 225 Writers		
	Top1	Top10	EER	Top1	Top10	EER
f1	36	74	7.23	30	77	10.97
f2	34	76	6.89	32	70	13.10
f3	42	81	6.56	34	72	14.34
f'4	43	77	6.96	40	75	12.74
f5	77	93	3.86	75	95	6.76
f6	46	83	7.11	41	76	12.70
f7	42	79	7.95	36	73	14.22

We first present the performance of the individual features (Table 2: numbers represent percentages) detailed in the above sections. Although the performance of the features varies significantly, it can be noticed that, for a chosen feature, the performance is more or less consistent across the two data sets. It is also evident that the local versions (*f5-f7*) of the three histograms (*f1-f3*) outperform their global counter parts, with ***f5*** (Stroke Direction Histogram) achieving the best results both on identification and verification tasks.

Table 3 Writer Recognition on Feature Combinations

Data Set / Feature	IAM 650 Writers			RIMES 225 Writers		
	Top1	Top10	EER	Top1	Top10	EER
f1-f3 (Global)	64	88	5.51	47	87	9.30
f1-f4 (Global)	78	93	3.51	63	87	9.05
f5-f7 (Local)	81	95	3.76	68	91	8.64
f1,f3,,f4,f5,f6	84	96	3.52	75	92	6.86
f3,f4,f5,f6	86	97	3.34	79	93	6.23

We then combine the features by computing the distance between two writings as an average (weighted and non-weighted) of the distances between the individual features, the weights being assigned relative to the performance of individual features. Table 3 summarizes some of the combinations that we tested. As with individual features, the histograms extracted from the local stroke information achieve better results. For writer identification, the highest rate we achieve stands at 86% for the IAM and 79% for the RIMES database. On the task of verification, we achieve EER of 3.34% and 6.23% on the two data sets respectively. The ROC curves for some of the feature combinations have been illustrated in figure 5.

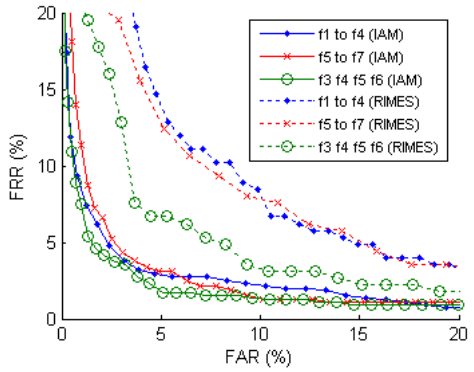

Figure 5. ROC Curves on the two data sets

Comparing the recognition performance across the two data sets, it can be seen that the identification and verification results are much better on the IAM writings than those on the RIMES both with individual features as well as with their combinations. This comes as no surprise as the RIMES data set is much more challenging than IAM; two samples of the same writer

might come from totally different writing conditions with a variable delay (at least few days) between the two collections.

Since the IAM database has been widely used in evaluating writer identification task, it would be interesting to present a comparative overview of the proposed methods. Table 4 summarizes the performance of the most recent studies on writer identification on this data set. Bulacu and Schomaker [5] currently hold the best performance results reading 89% on 650 writers. We have achieved an identification rate of 86% employing the proposed features and we hope to improve the results by adding new features and optimizing their selection.

Table 4 Comparison of writer identification methods

		Writers	Samples /writer	Performance
Marti et al. (2001)	[8]	20	5	90.7%
Bensefia et al. (2004)	[2]	150	2	86%
Schlapbach and Bunke (2006)	[12]	100	5/4	98.46%
Bulacu and Schomaker (2007)	[5]	650	2	89%
Our method		650	2	86%

5. Conclusion

We have presented here an effective method for writer recognition in handwritten documents. The method is based on finding the contours of a handwritten image and extracting a set of chain code based histograms at the global as well as local levels. The proposed features correspond to the characteristics that humans are very sensitive to. Also they are simple to compute and are very effective, realizing promising results on writer identification and verification. The proposed method is quite generic and can be applied to non-Latin languages such as Asian or Arabic scripts as well. In our future research, we intend to introduce additional contour based features as well as a feature selection mechanism which is likely to enhance the performance of our system.

6. References

[1] A. Bandera , C. Urdiales, F. Arrebola, F. Sandoval. 2D object recognition based on curvature functions obtained from local histograms of the contour chain code. Pattern Recognition Letters, 1999(20): pp.49-55.

[2] A. Bensefia, T. Paquet and L. Heutte: A writer identification and verification system, Pattern Recognition Letters, Vol 26, issue 13, 2005, pp. 2080-2092.

[3] V. Bouletreau, N. Vincent, R. Sabourin, H. Emptoz: Handwriting and signature:one or two personality identifiers?, In Proc.of 14th International Conference on Pattern Recognition, Los Alamitos, CA, (1998) 1758-1760.

[4] M. Bulacu, L. Schomaker, and L. Vuurpijl: Writer identification using edge-based directional features, In Proc. of 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, (2003) 937–941

[5] M. Bulacu and L. Schomaker: Text-Independent Writer Identification and Verification Using Textural and Allographic Features, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, pp. 701-717.

[6] E. Grosicki, M. Carré, J-M. Brodin, E. Geoffrois, "RIMES evaluation campaign for handwritten mail processing", In Proc of 11th Int'l Conference on Frontiers in Handwriting Recognition, Montreal, Canada, August 2008.

[7] U. Marti and H. Bunke, "A full English sentence database for off-line handwriting recognition", In Proc. of 5th International Conference on Document Analysis and Recognition, Bangalore, India, 1999, pp. 705-708.

[8] U. Marti, R. Messerli, H. Bunke, "Writer Identification Using Text Line Based Features", 6th Int'l Conference on Document Analysis and Recognition, 2001. icdar, p. 0101.

[9] A. Nosary, L. Heutte, T. Paquet, Y. Lecourtier, "Defining writer's invariants to adapt the recognition task", In Proc. of 5th International Conference on Document Analysis and Recognition, Bangalore, India, (1999) 765-768

[10] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification – the state of the art", Pattern Recognition, vol. 22, n°2, (1989) 107-131

[11] H.E.S. Said, T.N Tan, K.D. Baker, "Personal Identification Based on Handwriting", Pattern Recognition, vol. 33, (2000) 149-160.

[12] A. Schlapbach and H. Bunke, "Off-line Writer Identification Using Gaussian Mixture Models". In Proc. of 18th Int. Conf. on Pattern Recognition, pages 992–995. August 2006, Hong Kong.

[13] I. Siddiqi and N. Vincent, "Combining Global and Local Features for Writer Identification", In Proc of 11th Int'l Conference on Frontiers in Handwriting Recognition (ICFHR), 2008,Canada.

[14] S. Srihari, S. Cha, H. Arora, and S. Lee, "Individuality of handwriting", J. of Forensic Sciences, 47(4):1.17, (2002)

[15] E.N. Zois, and V. Anastassopoulos, Morphological waveform coding for writer identification. Pattern Recognition 2000, 33 (3), 385–398.