Writer Recognition by Combining Local and Global Methods

Karl-Heinz Steinke, Martin Gehrke, Robert Dzido
University of Applied Sciences and Arts, Hanover
Hanover, Germany
karl-heinz.steinke@fh-hannover.de

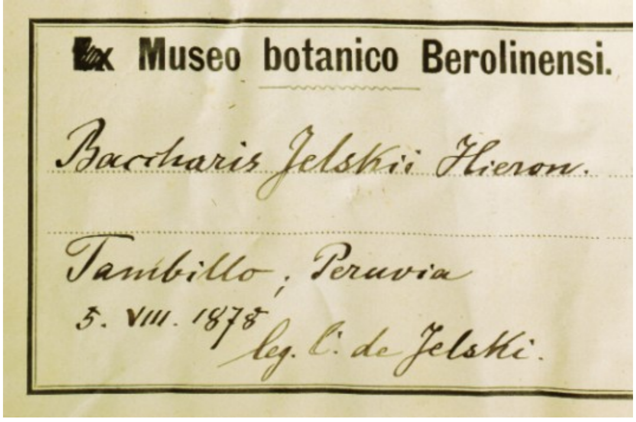
Abstract*— The research project „Herbar Digital" was started in 2007 with the aim to digitize 3.5 million dried plants on paper sheets belonging to the Botanic Museum Berlin in Germany. Frequently the collector of the plant is unknown, so a procedure had to be developed in order to determine the writer of the handwriting on the sheet. In the present work the static character was transformed into a dynamic form. This was done with the model of an inert ball which was rolled along the written character. During this off-line writer recognition, different mathematical procedures were used such as the reproduction of the write line of individual characters by Legendre polynomials. When only one character was used, a recognition rate of about 40% was obtained. By combining multiple characters, the recognition rate rose considerably and reached 98.7% with 13 characters and 93 writers (chosen randomly from the international IAM-database [3]). A global statistical approach using the whole handwritten text resulted in a similar recognition rate. By combining local and global methods, a recognition rate of 99.5% was achieved.

***Keywords*- handwriting, writer recognition, forensics**

I. Introduction

On a herbarium sheet many facts are noted, such as the site where it was found (if possible GPS coordinates), and the date it was found, collector and usually the scientific name of the plant. Among the approximately 3.5 million plants, there are many whose collector is unknown, because an appropriate note is missing on the herbarium sheet. An example of a label with handwritten data is shown in figure 1. These sheets should be allocated by the analysis of the handwriting contained in it. Also in other fields, e.g. in forensics, the allocation of handwriting to the corresponding author plays an important role. In literature there are many approaches to writer recognition which are mainly based on data of a digitization tablet. There, the handwriting coordinates are collected in real time and so the coordinate sequence can be evaluated. But in our case we are dealing with old static handwritings of writers who are no longer present. This so-called off-line writer recognition represents a more complicated problem, since no coordinate sequences are available.

Research in automatic identification of writers focused mainly on the statistical approach. This led to the extraction of characteristics such as run lengths [10] and inclination distributions as well as entropy characteristics. Newer approaches, e.g. that of Siddiqi [12] try to combine global and local features, but still with modest success. Niels [5] uses character prototypes and differentiates writers on the basis of how often the prototypes occur in a long text. For this, a time-consuming analysis of the characters has to be made by a handwriting expert. Srihari [6] developed individuality-characteristics for static pictures by extraction of macro and micro features. It was shown that individual characters possess different capabilities of discriminating between writers. Said [11] presents a global approach and regards the handwriting as different textures, which he received by application of the Gabor filters and the co-occurrence matrix. Marti [4] analyzes the difference in handwritings by structural characteristics of each text line. Schomaker [8] uses the contour of connected components. Bensefia [1] uses local characteristics which originate from the analysis of the upper contour's minima.



Ex Museo botanico Berolinensi.

Baccharis Jelskii Hieron.

Tambillo, Peruvia

5. VIII. 1878 leg. C. de Jelski.

Figure 1. Label with handwriting

In order to achieve a higher recognition rate, a new approach was developed in this work which transfers static handwriting into dynamic coordinate sequences. By the subsequent treatment of these x and y-coordinates with different algorithms, characteristics are obtained from single characters, which suggest the possible writer of the characters.

II. Converting Static Characters into Dynamic Sequences

During on-line writing recognition, e.g. with PDAs, time series are used to read handwriting. In order to make a similar approach possible for already written handwriting on paper, a software was developed which transfers handwriting into dynamic coordinate sequences. The handwritten character can be extracted out of a connected text. In order to recover the write line from old documents, whose writer is no longer present the following model serves: The writing is written as a groove in sand. A ball equipped with inertia rolls is rolled along the groove and reproduces the write line while keeping its last direction. If it arrives at a terminator point, it will run back and try to deviate from the last way. In unclear situations the ball can be pulled with a „rubber band " (right mouse button) in the desired direction. If characters merge with the next text line, the write line (see figure 2) will be received as well.

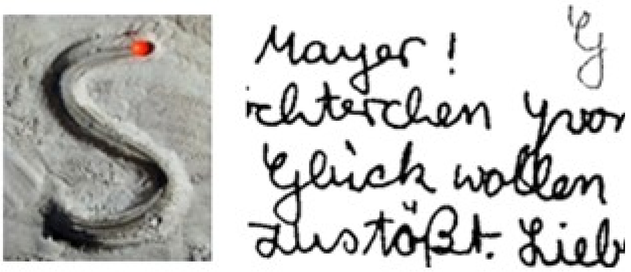


Figure 2. Ball and handwriting with extracted G

The dynamic coordinate sequence is to a large extent independent of the used writing pen. In figure 3 a "W" from a specimen of handwriting and the coordinate sequence are compared.

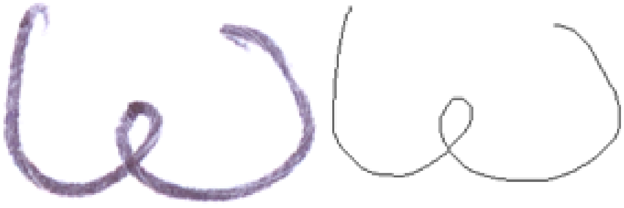


Figure 3. Comparison original and sequence of W

III. Local Method

In order to compare characters and writers, the following procedures were explored:

-Reproduction by vectors

-Approximation by Fourier series

-Approximation by Chebychev polynomials

-Comparison by cross correlation

-Approximation by Legendre polynomials

-Comparison by image moments

A. Vectors

With the vector comparison the coordinate sequences of the individual characters are converted into angles. They are added according to the respective vector length and computed from the middle angles. These angles can be compared with those of the other existing characters. The differences received thereby give information on the character's similarity. The determination of the angles takes place according to a system which is similar to the Freeman code [2]. In figure 4 the pattern of the coding is represented. For illustration, a "C" and its reproduction are represented by 4 vectors.

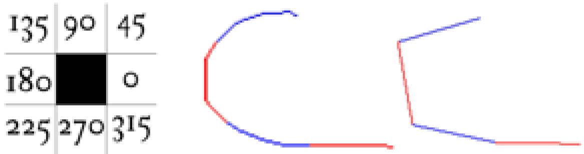


Figure 4. Angle code, original C and reproduction

B. Fourier series

By a Fourier expansion a repetitive function can be represented as a set of sine and cosine functions, whose frequencies are integral multiples of the basic frequency ω=2π/T.

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cdot \cos(n\omega t) + b_n \cdot \sin(n\omega t))$$

The Fourier coefficients a_n und b_n can be computed by the Euler formulas:

$$a_n = \frac{2}{T}\int_c^{c+T} f(t)\cos(n\omega t)dt$$

$$b_n = \frac{2}{T}\int_c^{c+T} f(t)\sin(n\omega t)dt$$

f(t) is approximated by finite trigonometric polynomial $f_n(t)$.

$$f_n(t) = \frac{a_0}{2} + \sum_{k=1}^{n}(a_k \cdot \cos(k\omega t) + b_k \cdot \sin(k\omega t))$$

By the coefficients the character can be back-transformed. In figure 5, an original "C" and an approximation with 64 coefficients is shown.

Figure 5. Original C and Fourier reproduction

C. *Chebychev polynomials*

Functions can be approximated by the use of Chebychev polynomials of the first kind with a very high accuracy. The polynomials are computed by the following formula:

$$T_n(x) = \cos(n \cdot ar\cos(x)), x \in [-1,1]$$

A polynomial $T_n(x)$ has exactly n zeros in the interval [-1,1]. In figure 6 the Chebychev polynomials up to the 4th order are represented.

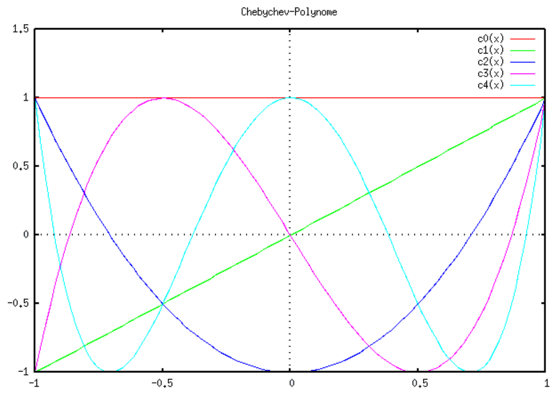


Figure 6. Chebychev polynomials

In order to get the Chebychev coefficients needed in this application, it is necessary to approximate the writing process of the characters by the following expression:

$$f(x) \approx \sum_{k=o}^{N-1} c_k T_k(x) - \frac{1}{2} c_0$$

N=number of coefficients

On the basis of the differences between the coefficients, the most similar characters can be determined.

D. *Cross correlation*

With this method, the number of points of the characters is reduced by a sub sampling process. An example of a "C" reduced to 32 points is represented in figure 7.

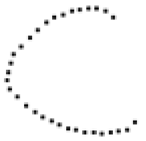


Figure 7. Sub sampling with 32 points

For each character the average value of the expansion in x and y-direction is calculated and divided by the number of points of n of the character

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

In order to reach a scaling invariance, a standardization of the reduced characters is necessary. The standard deviation of the expansion in x and y-direction is computed.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

The individual points of the character are compared in each case with those of the other existing characters. By the sum of the distances, information on the characters with the smallest deviation is obtained. Figure 8 shows the result of a comparison. From left to right, a “C”, the character with the largest similarity and the one between both are shown.

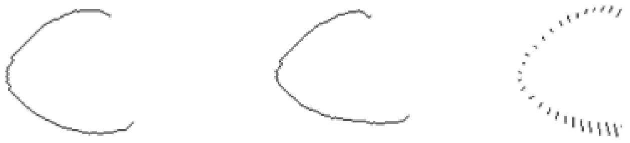


Figure 8. C, most similar character, differences

E. *Legendre polynomials*

In the 5th method for the determination of the character with the smallest deviation, the Legendre coefficients are used. In connection with the pertinent Legendre polynomials, functions can be approximated. The ones generally represented look as follows:

$$P_n(x) = \frac{1}{(2^n n!)} \cdot \frac{d^n}{dx^n}\left[(x^2 - 1)^n\right]$$

A polynomial $P_n(x)$ has exactly n zeros in the interval [-1,1] and between two zeros of $P_n(x)$, there exists one zero of $P_{n+1}(x)$. An approximation of the functions by the polynomials is done by

$$f(x) = \sum_{n=1}^{\infty} c_n P_n(x)$$

where c_n are the Legendre coefficients. They can be computed by

$$c_n = \frac{2n+1}{2} \int_{-1}^{1} f(x) P_n(x) dx$$

After the calculation a back transformed function can be produced by inserting the coefficients and the polynomials into the expression. An example of an approximation by the discrete Legendre transformation is given in figure 9 where "C" as original and as inverse transform is shown.

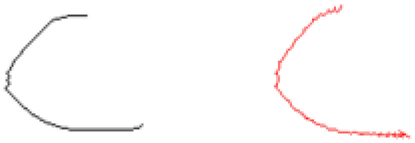


Figure 9. Original C, back transformed C

F. Image moments

The discrete geometrical moments are defined by the following expression:

$$m_{p,q} = \iint x^p y^q f(x,y) dxdy$$

Eccentricity ε is the deviation from the roundness of an object. With a circle the value for ε is 0 and with a straight line the value is 1. For all other objects the values lie between 0 and 1. "C" represented in figure 10 has a value 0.0719, while the "I" shown has an eccentricity of 0.9931.

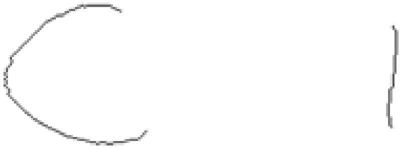


Figure 10. Low eccentricity, high eccentricity

G. Splitting characters into two time series

By the Fourier series, the Chebychev and the Legendre polynomials functions can be approximated. As you can see in figure 11 the character "Z" is not a function. Up to three y-values for one x-value are existing. Therefore an approximation of the writing process cannot happen directly by one function. The problem is solved as the write line is split up into x and y-movement and afterwards each is separately discretely approximated.

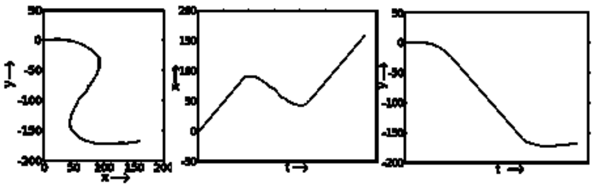


Figure 11. x-y-diagram, x-t-diagram, y-t-diagram

H. Reconstruction of characters

With the help of the characteristics of most used comparison methods the original writing process can be reconstructed again. The accuracy of these reproductions depends on the number of used features. The figures 12 and 13 show the reproductions of a "W" by vectors, Fourier series, sub sampling method and Legendre polynomials. With all these methods the represented reproductions are numbered as follows:

Original character(1) , reconstruction by 64 features(2), 32 features(3), 16 features(4), 8 features(5), 4 features(6).

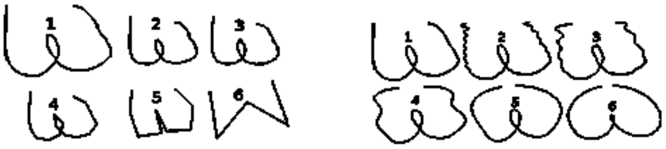


Figure 12. Reconstruction by vectors and Fourier series

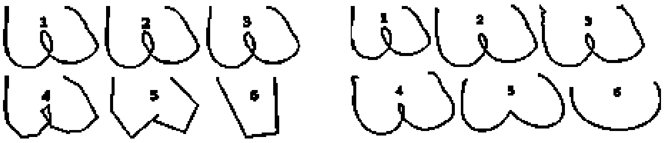
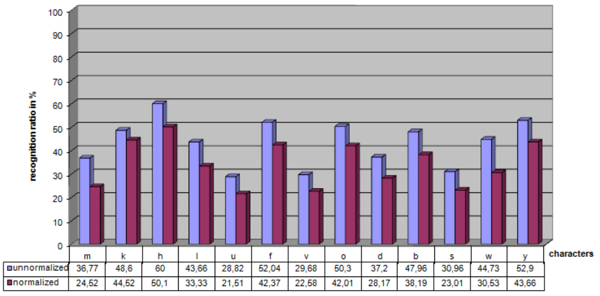


Figure 13. Reconstruction by sub sampling and Legendre polynomials

I. Results

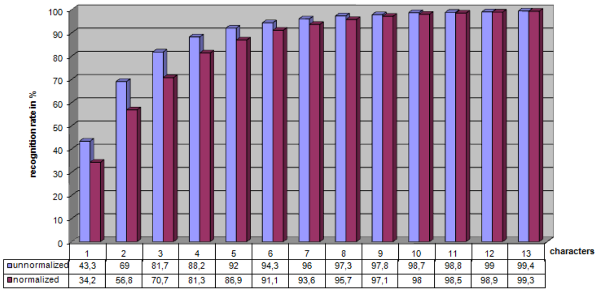
For the comparison with a large number of writers an international data base from Switzerland was chosen. The IAM-database [3] contains handwritings in English language with different texts. 6045 characters from 93 writers were extracted from the images (5 samples randomly chosen from each writer). The 13 characters m, k, h, l, u, f, v, o, d, b, s, w, y were selected (see figure 16). The characters possess different discriminatory abilities. In figure 14 the writer recognition rates using only one character is shown.



	m	k	h	l	u	f	v	o	d	b	s	w	y
unnormalized	36,77	48,6	60	43,66	28,82	52,04	29,68	50,3	37,2	47,96	30,96	44,73	52,9
normalized	24,52	44,52	50,1	33,33	21,51	42,37	22,58	42,01	28,17	38,19	23,01	30,53	43,66

Figure 14. Recognition rates with one character

The combination of several characters promises a higher recognition rate by more information. For the comparison of the writers n characters of an unknown writer are selected. To each of the selected characters the most similar character of any of the well-known writers is assigned. The distances of the n most similar ones are added and a decision is made in favor of the writer with the minimum distance sum. It arises in figure 15 that the recognition rate with a higher number of used characters rises considerably. When combining the characters, the procedures with the Legendre, Fourier and Chebychev coefficients prove themselves as almost equivalent. It is interesting that the size-normalized characters first supply worse recognition rates. With increasing character number they measure up with the not normalized characters.



	1	2	3	4	5	6	7	8	9	10	11	12	13
unnormalized	43,3	69	81,7	88,2	92	94,3	96	97,3	97,8	98,7	98,8	99	99,4
normalized	34,2	56,8	70,7	81,3	86,9	91,1	93,6	95,7	97,1	98	98,5	98,9	99,3

Figure 15. Recognition rates with multiple characters

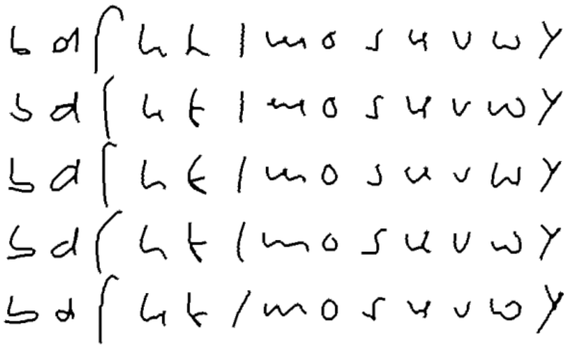


Figure 16. Thirteen characters of one writer

IV. GLOBAL METHOD

In order to analyze the whole handwriting of different writers, a statistical approach [10] was selected. The handwriting is seen as a texture with a steady structure of line elements all over the image. For the description of such a texture, a suitable set of primitive elements has to be found whose frequency of occurrence is suited to distinguishing different writers to the greatest possible extent. The line segments of which the writing is composed can be taken as primitive elements of a handwriting specimen. Straight line segments may be obtained by the run lengths of pixel chains. The number and length of pixel chains is determined in eight different directions (see figure 17) and for each direction a frequency distribution is made. The features obtained by this shift-invariant transformation are nearly text independent, as long as there is enough text at hand (about three to five handwriting lines). The feature vector furnishes information about the sloping position, size, regularity and roundness of the handwriting. The developed software can be imagined as a shredder (see figure 17). The feature vectors obtained by the described method have a very high dimension. As neighbored components of the vector are strongly correlated, they are added to a certain degree so that only 8 features in each direction remain. Altogether a feature vector with 64 components results (see figure 18).

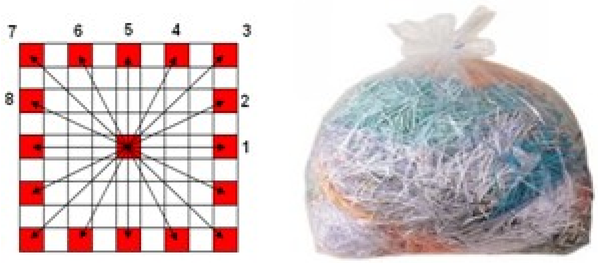


Figure 17. Eight directions for the shredder

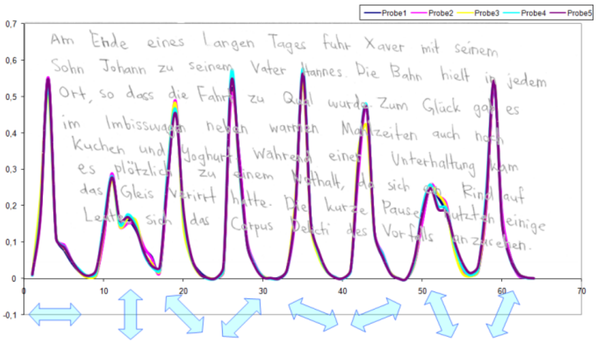


Figure 18. Feature vectors of 5 samples of 1 writer

V. COMBINING LOCAL AND GLOBAL METHOD

From the IAM-database [3] writers were chosen with at least 5 samples of each. 93 writers were found and so 93*5=465 sheets with handwritings were processed. From each sheet 13 characters were extracted with the local method described in section 2. With a simple nearest neighbor classifier and the leaving one out method the correct writer was found in 459 cases. Only 6 samples [44(5),48(1),50(2),50(5),75(5),84(1)] were assigned to a wrong writer. The global method described in section 4 also mismatched only 6 but disjoint samples [25(5),32(3),37(2),40(1),52(4),88(2)]. Local and global methods were combined by computing the positions in both hit lists. The decision is made in favor of the writer with the minimum sum of the two hit lists positions. The error rate falls to 2 of 465 samples (see figure19).

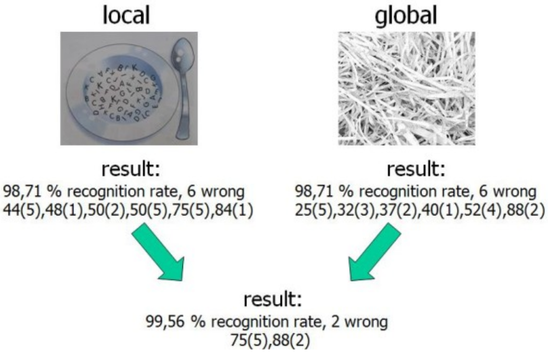


Figure 19. Combining local and global method

ACKNOWLEDGMENT

This project is financed by the state of Lower Saxonia and the Volkswagen Foundation.

REFERENCES

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

[2] Freeman, H., on the encoding of arbitrary geometric configurations, I R E Trans. EC-10, 260-265 (1961)

[3] Marti, U., Bunke, H., A full english sentence database for off-line handwriting recognition, Proceedings of the 5. Int. Conference on Document Analysis and Recognition, Bangalore 1999, pp. 765-768.

[4] U.V. Marti, R. Messerli, H. Bunke, Writer Identification Using Text Line Based Features, Proc. Of the 6th International Conference on Document Analysis and Recognition, Seattle, USA, 2001, pp. 101-105.

[5] Niels, R., Grootjen, F., Vuurpijl, L., writer identification through information retrieval: the allograph weight vector, Proceedings of the 11. Int. Conference on Frontiers in Handwriting Recognition, Montreal, 2008.

[6] S. Srihari, S. , H. Arora, S. Lee, Individuality of handwriting, J. of Forensic Sciences, 47(4):1.17, July 2002

[7] Schlapbach, Andreas; Bunke, Horst: Off-line Handwriting Identification Using HMM Based Recognizers, 2004, publications Uni Bern

[8] L. Schomaker, L. And M. Bulacu, Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script, in IEEE Transactions of Pattern Analysis and Machine Intelligence, vol. 26, no. 6, pp. 787-798, 2004.

[9] Steinke, K.-H., Dzido, R., Gehrke, M., Prätel, K., Feature recognition for herbarium specimens (Herbar-Digital), Proceedings of TDWG, Perth, 2008

[10] Steinke, K.-H., Recognition of Writers by Handwriting Images; Conference on Pattern Recognition, 1980, Oxford, published in Pattern Recognition 1981; M. Duff Ed.

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

[12] Siddiqi, I., Vincent, N., Combining global and local features for writer identification, Proceedings of the 11. Int. Conference on Frontiers in Handwriting Recognition, Montreal, 2008