Writer Recognition by Characters, Words and Sentences

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

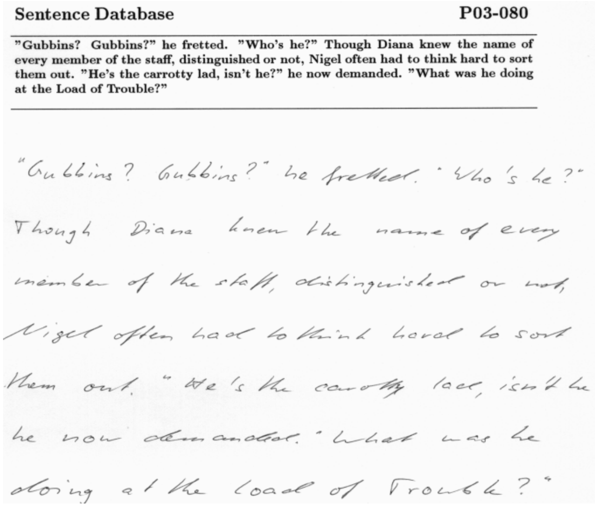
***Abstract*— The methods developed in the research project „Herbar Digital" are to help plant taxonomists to master the great amount of material of about 3.5 million dried plants on paper sheets belonging to the Botanic Museum Berlin in Germany. Frequently the collector of the plant is unknown. So a procedure had to be developed in order to determine the writer of the handwriting on the sheet. In the present work the static character is transformed into a dynamic form. This is done with the model of an inert ball which is rolled through the written character. During this off-line writer recognition, different mathematical procedures are used such as the reproduction of the write line of individual characters by Legendre polynomials. When only one character is used, a recognition rate of about 40% is obtained. By combining multiple characters, the recognition rate rises considerably and reaches 98.7% with 13 characters and 93 writers (chosen randomly from the international IAM-database [3]). Another approach tries to identify the writer by handwritten words. The word is cut out and transformed into a 6-dimensional time series and compared e.g. by means of DTW-methods. A global statistical approach using the whole handwritten sentences results in a similar recognition rate of more than 98%. By combining the methods, a recognition rate of 99.5% is achieved.**

Keywords- handwriting, writer recognition, forensics

I. Introduction

Among the approximately 3.5 million herbarium sheets, there are many whose collector is unknown, because an appropriate note is missing. These sheets should be allocated by the analysis of the handwriting contained in it. Also in forensics, the allocation of handwriting such as threatening letters and forged cheques 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 not 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.



Sentence Database **P03-080**

"Gubbins? Gubbins?" he fretted. "Who's he?" Though Diana knew the name of every member of the staff, distinguished or not, Nigel often had to think hard to sort them out. "He's the carrotty lad, isn't he?" he now demanded. "What was he doing at the Load of Trouble?"

"Gubbins? Gubbins?" he fretted. "Who's he?"
Though Diana knew the name of every
member of the staff, distinguished or not,
Nigel often had to think hard to sort
them out. "He's the carotty lad, isn't he
he now demanded. "What was he
doing at the Load of Trouble?"

Figure 1. Handwriting from IAM-database

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. Characters Method

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 documents, whose writer is not present the following model serves: The writing is written as a groove in sand. A ball equipped with inertia rolls 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.

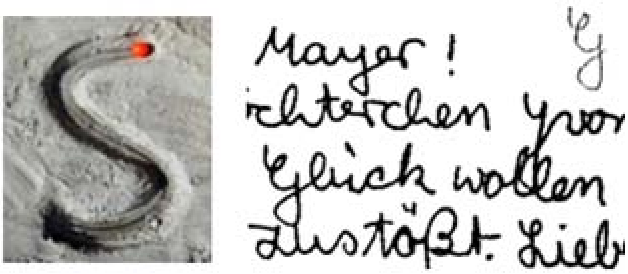


Figure 2. Ball and handwriting with extracted G

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.

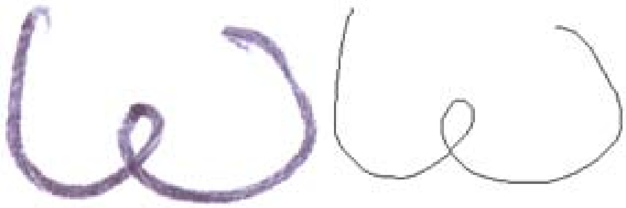


Figure 3. Comparison original and sequence of W

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

-Approximation by Fourier series

-Approximation by Chebychev polynomials

-Approximation by Legendre polynomials

A. 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 4, an original “C” and an approximation with 64 coefficients is shown.

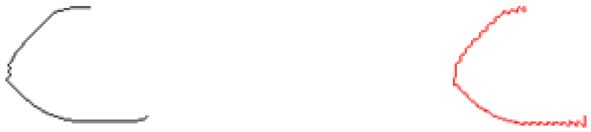


Figure 4. Original C, back transformed C

B. 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 5 the Chebychev polynomials up to the 4th order are represented.

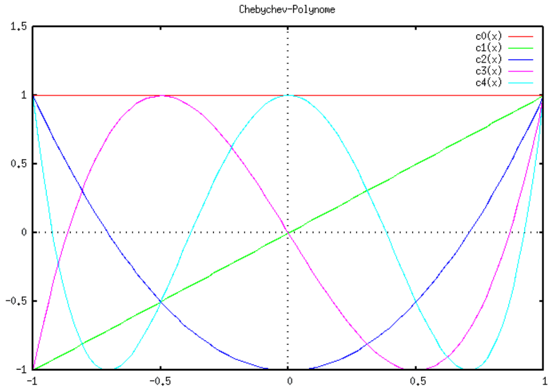


Figure 5. Chebychev polynomials

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.

C. *Legendre polynomials*

In the 3th 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 6 where "C" as original and as inverse transform is shown.

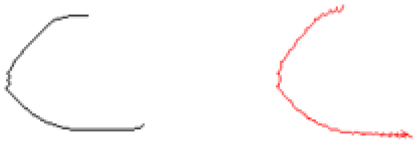


Figure 6. Original C, back transformed C

D. *Splitting characters into time series*

By the Fourier series, the Chebychev and the Legendre polynomials functions can be approximated. As you can see in figure 7 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.

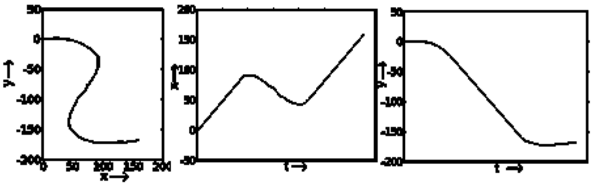


Figure 7. x-y-diagram, x-t-diagram, y-t-diagram

E. *Reconstruction of characters*

With the help of the characteristics of all used comparison methods the original writing process can be reconstructed again. The accuracy of these reproductions depends on the number of used features. Figure 8 shows the reproductions of a "W" by Legendre polynomials and Fourier series. With both 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).

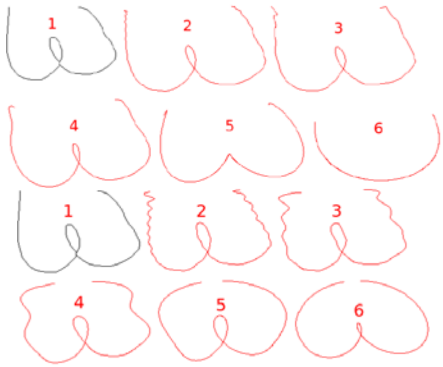


Figure 8. Reconstruction by Legendre polynomials and Fourier series

F. 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 11). The characters possess different discriminatory abilities. In figure 9 the writer recognition rates using only one character is shown.

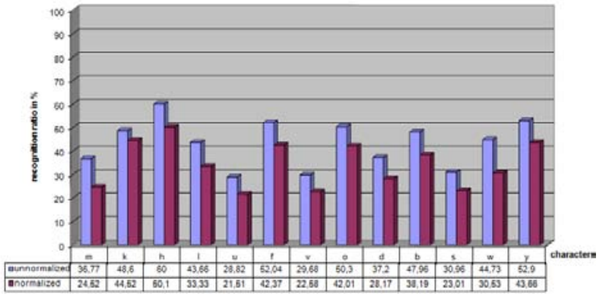


Figure 9. Recognition rates with one character

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 10 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.

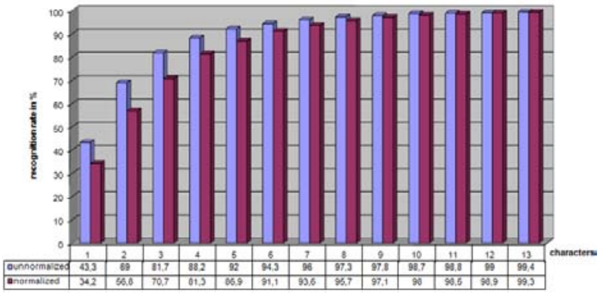


Figure 10. Recognition rates with multiple character

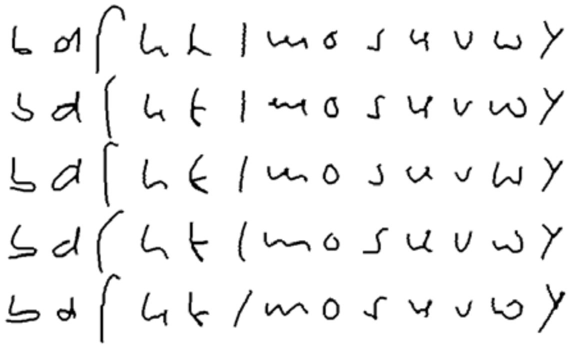
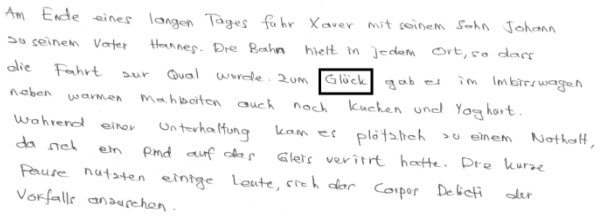


Figure 11. Thirteen characters of one writer

III. WORDS METHOD

For writer recognition by words a multiple occurrence of a word is needed. Unfortunately the IAM-database does not contain enough words. So an own german database with a unique text (see figure 12) had to be established.



Am Ende eines langen Tages fuhr Xaver mit seinem Sohn Johann
zu seinem Vater Hannes. Die Bahn hielt in jedem Ort, so dass
die Fahrt zur Qual wurde. Zum Glück gab es im Imbisswagen
neben warmen Mahlzeiten auch noch Kuchen und Yoghurt.
Wahrend einer Unterhaltung kam es plötzlich zu einem Nothalt,
da sich ein Rind auf das Gleis verirrt hatte. Die kurze
Pause nutzten einige Leute, sich das Corpus Delicti des
Vorfalls anzusehen.

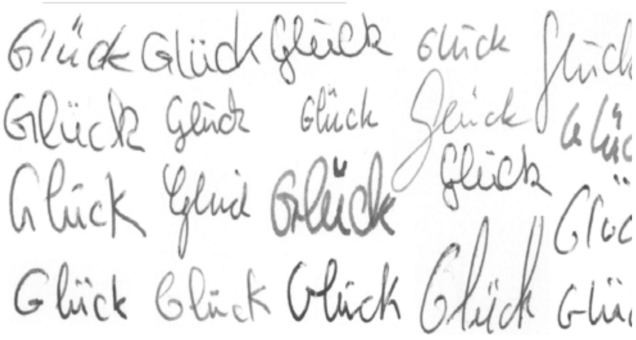


Figure 12. Handwriting of the german database

Figure 13. Word “Glück” of different writers

For approximation by Fourier series, Chebychev and Legendre polynomials functions had to be constructed from the words. As you can see in figure 14 neither the original word “Glück” nor the word with slant and slope correction is a function.

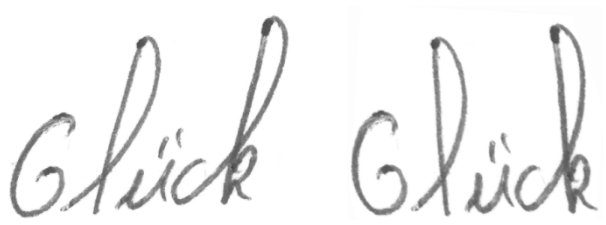


Figure 14. Original word and word with slant and slope correction

A. Splitting words into six time series

The problem is solved as the word is split up into 6 time series (see figure 19) and afterwards each is separately discretely approximated. The 6 time series are: upper writing line (yellow), lower writing line (turquoise), mass (green), hollow space (blue), gravity (white) and variance (red). The possible steps are shown in figure 15-17. Slant and slope correction proved as valuable in word recognition. In writer recognition step 3 and 4 are skipped. After mirroring the word along the main axis (see figure 18) the procedure is repeated.

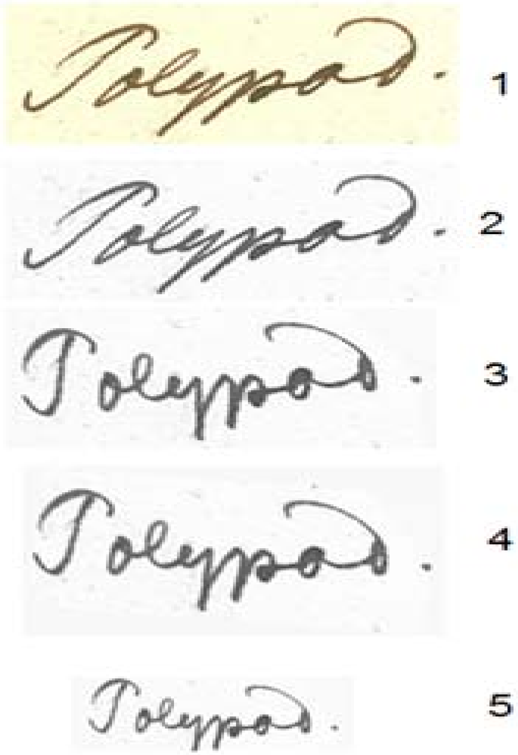


Figure 15. 1 Original word, 2 Gray level image, 3 Slant correction, 4 Slope correction, 5 Normalization

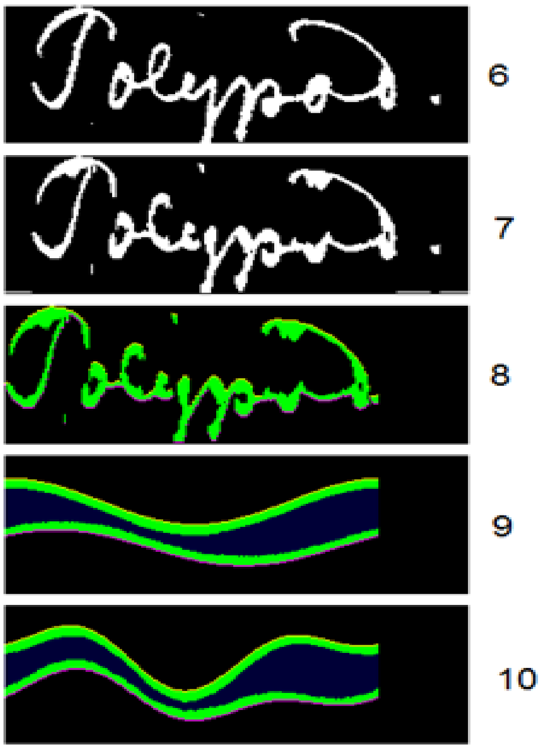


Figure 16. 6 Binary image, 7 Reconstructed word by time series, 8 Holes removed, 9 Reconstructed by 2 Coefficients, 10 Reconstructed by 4 Coefficients

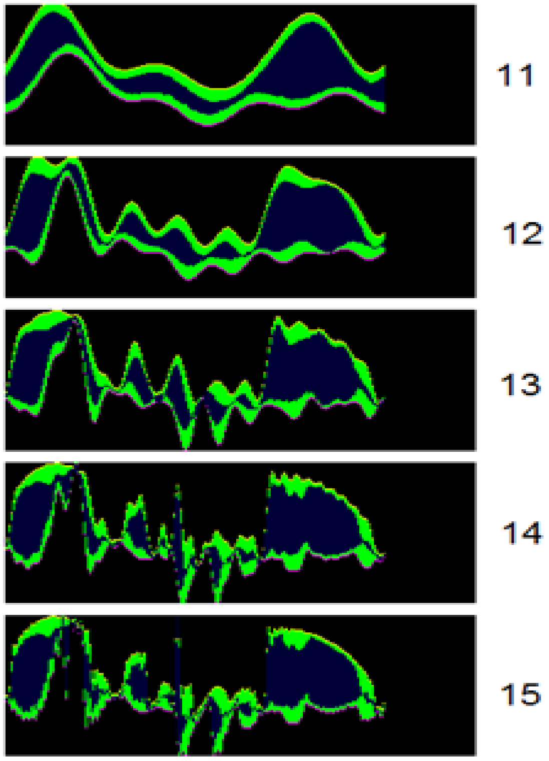


Figure 17. 11 Reconstructed by 8 Coefficients, 12 Reconstructed by 16 Coefficients, 13 Reconstructed by 32 Coefficients, 14 Reconstructed by 64 Coefficients, 15 Reconstructed by 128 Coefficients

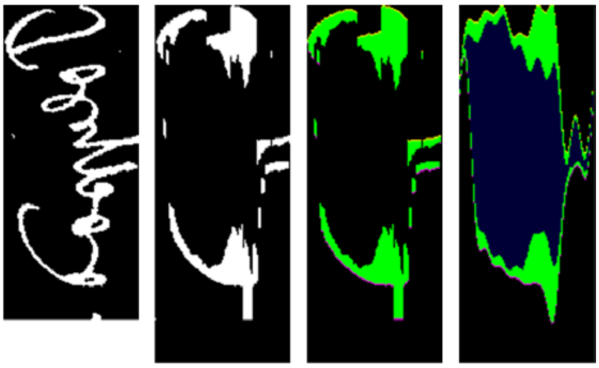


Figure 18. Mirroring the word along the main axis

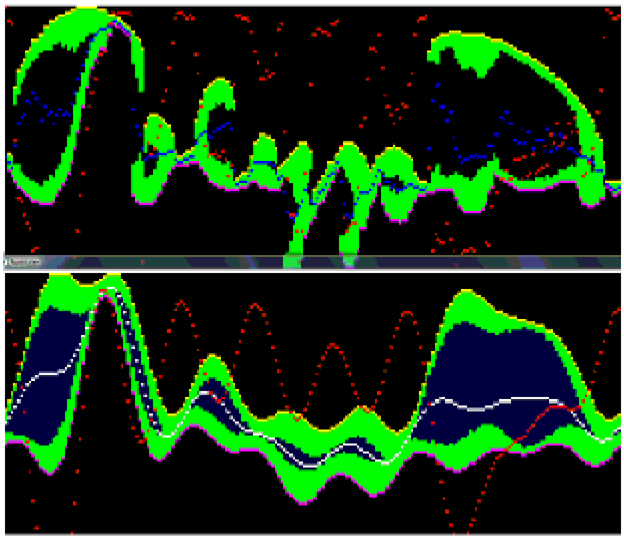


Figure 19. 6 time series and low pass filtering

B. *Results*

For the comparison with a large number of writers the german data base was chosen. 6240 words from 104 writers were extracted from the images. The words possess different discriminatory abilities. In figure 21 the writer recognition rates using only one word is shown. The results with dynamic time warping method (see figure 20) are a bit better but the method is very time consuming.

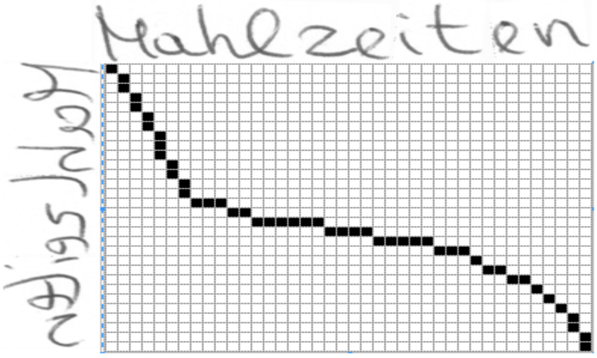
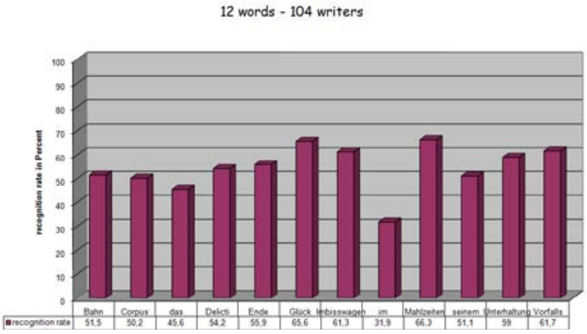


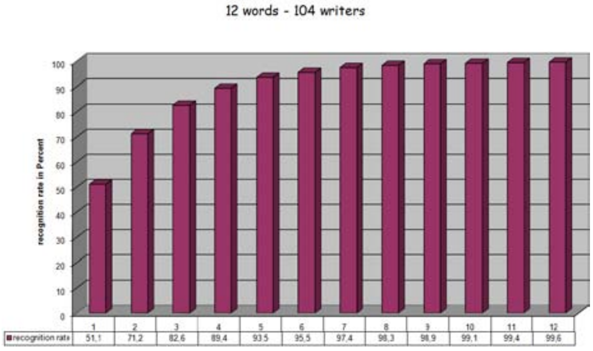
Figure 20. Dynamic time warping



	Bahn	Corpus	das	Delicti	Ende	Glück	Imbisswagen	im	Mahlzeiten	seinem	Unterhaltung	Vorfalls
■ recognition rate	51,5	50,2	45,6	54,2	55,9	65,6	61,3	31,9	66,3	51,1	59	61,7

Figure 21. Recognition rates with one word

The combination of several words promises a higher recognition rate by more information. It arises in figure 22 that the recognition rate with a higher number of used words rises considerably. When combining the words, the procedures with the Legendre, Fourier and Chebychev coefficients prove themselves as almost equivalent.



	1	2	3	4	5	6	7	8	9	10	11	12
■ recognition rate	51,1	71,2	82,6	89,4	93,5	95,5	97,4	98,3	98,9	99,1	99,4	99,6

Figure 22. Recognition rates with multiple words

IV. SENTENCES METHOD

In order to analyze the whole handwriting sentences 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 23) 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 23). 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 24).

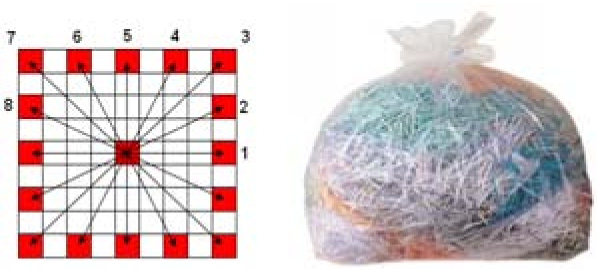


Figure 23. Eight directions for the shredder

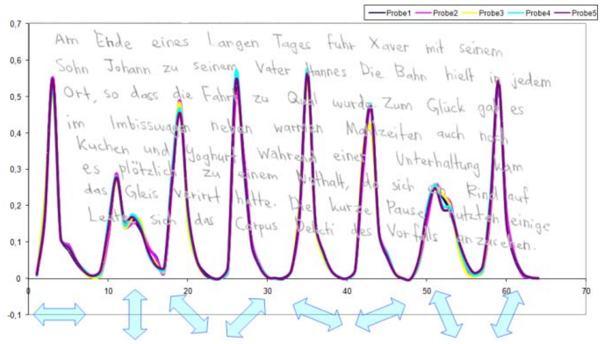


Figure 24. Feature vectors of 5 samples of 1 writer

V. Combining characters and sentences 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 sentences method described in section 4 also mismatched only 6 but disjoint samples [25(5), 32(3), 37(2), 40(1), 52(4), 88(2)].

Characters and sentences 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 figure 25).

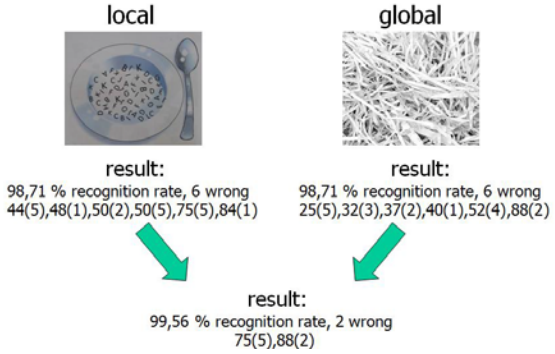


Figure 25. Combining characters and sentences method

VI. Combining characters and words method

From the german database 104 writers (see figure 27) were taken with 5 samples of each. From the 520 sheets with handwritings 11960 capital letters were extracted. With a simple nearest neighbor classifier and the leaving one out method the correct writer was found in 518 cases. Only 2 samples [51(5), 90(2)] were assigned to a wrong writer.

The words method with 6240 words also mismatched only 2 but disjoint samples [32(1), 16(2)]. Characters and words 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 0 of 520 samples (see figure 26).

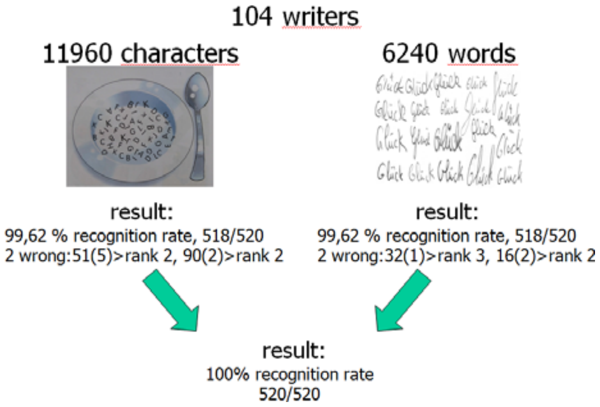

104 writers
11960 characters
6240 words
result:
99,62 % recognition rate, 518/520
2 wrong:51(5)>rank 2, 90(2)>rank 2
result:
99,62 % recognition rate, 518/520
2 wrong:32(1)>rank 3, 16(2)>rank 2
result:
100% recognition rate
520/520

Figure 26. Combining characters and words method

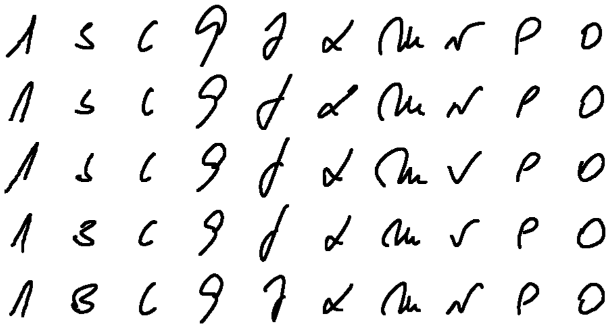


Figure 27. Capital letters of one writer

VII. Conclusions

Writer recognition by means of characters is a very efficient method especially when using multiple characters. The extraction of one character takes about one minute. The method can also be applied if there is not much text at hand. Also in combination with the global sentences method it leads to high recognition rates. Writer recognition by words is possible as shown in section 3. However it cannot be used for a recognition system in forensics, because the words needed for comparison will often not be available.

In the context of the project "Herbar Digital" a software prototype is going to be developed which can be brought to market later. It should be able to compare handwritten proofs and assign them to the respective author. Based on this prototype a software platform is to be developed, so that museums can gather and evaluate existing proofs with handwritten comments. A secondary use in forensics is possible. The software algorithm can be developed further and used for handwriting comparison in forensic investigations.

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

[13] Steinke, K.-H., Lokalisierung von Schrift in komplexer Umgebung, Tagungsband der Jahrestagung der deutschen Gesellschaft für Photogrammetrie, Jena März 2009