New Features for the Recognition of German-Kurrent-Handwriting with HMM-based Offline Systems

Klaus Prätel
Department of Computer Vision and Remote Sensing, TU Berlin, Marchstr. 23, 10587 Berlin, Germany
klaus.praetel@campus.tu-berlin.de

Keywords: New Geometrical Features, German Cursive.

Abstract: In 2007, the project Herbar-Digital was launched at the University of Applied Sciences and Arts Hannover (Steinke, K.-H., Dzido, R., Gehrke, M., Prätel, K., 2008). The aim of this project is to realize a global Herbarium, to compare findings quickly and catalog. There are many herbaria, i.e. collections of herbarium specimens worldwide. Herbarium specimens are paper pages where botanical elements are glued on. This herbarium specimens are provided with some important clues, such as the name of the submitter, barcode, color table, flag for first record, description of the findings, this often handwritten. All information on the herbarium specimens should be evaluated digitally. Since a number of discoveries in the 19th Century took place, Alexander von Humboldt is to be mentioned, here is the challenge to identify specific manuscripts from this period. This paper describes the topic recognition of old German handwriting (cursive).

1 INTRODUCTION

Approach to handwriting recognition:

Writer independent offline handwriting recognition.

Based on architecture of Hidden Markov Models.

Module for extracting text lines from image data.

Modules for pre-processing such as Slant, Slope, and Scale.

Sliding-Window Serialization.

Automatic limitation of the model states as a function of training material (30 as the upper limit).

Parameter-Estimation using Standard-Baum-Welch-Training.

Decoding (recognition) on the Viterbi algorithm.

Semi-continuous HMMs (state reduction) for Writing Model.

Statistical n-gram models as language models.
Linear and Bakis topology.

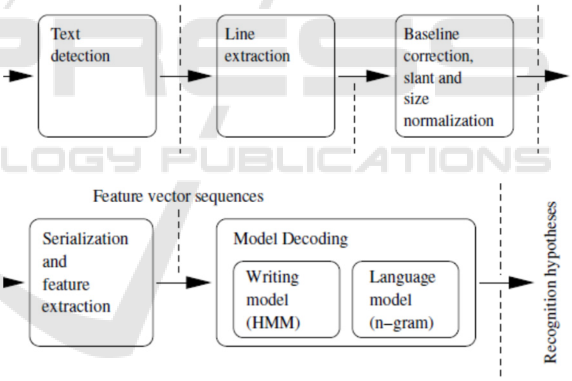


Figure 1: Schematic representation of a typical architecture of a handwriting recognition system (Fink, G. A., 2008).

Preprocessing:

- Binarization

  Separation of the foreground from the background.

- Skeletonization

  Thinning by line following, reduction to one-pixel thickness.

- Determination of reference lines

  Upper limit line, midline, baseline, lower profile

Prätel, K.
New Features for the Recognition of German-Kurrent-Handwriting with HMM-based Offline Systems.
DOI: 10.5220/0006185403660371
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 366-371
ISBN: 978-989-758-222-6

- Correction of font orientation

  Estimate of the font baseline

- Correction of font size

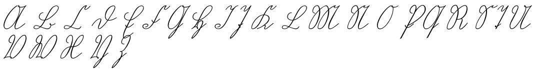
- Correction of the slant

- Feature extraction

Comparison of German cursive hand with the simplified output font:

As you can see the writing samples, a cursive hand, here just the large letters, is characterized by significantly increased crossing processes (possibly multiple entry between upper and lower contour) than in the Simplified output font.

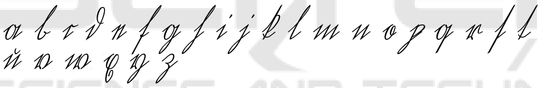
German cursive (uppercase)



Simplified output font (uppercase)

A B C D E F G H I J K L
M N O P Q R S T U V W X
Y Z

German cursive (lowercase)



Simplified output font (lowercase)

a b c d e f g h i j k
l m n o p q r s t u v
w x y z

Feature extraction

Now features are determined to make the following step, a classification can be used for the pre-segments. On one hand, the features may be discriminative, being very different from each other when they come from different classes. On the other hand, they should differ little when they come from the same class. In order to achieve this, to realize this, the peculiarities of the German cursive must be considered. Knowledge: the feature extraction must be extended. The following are the major principles of the feature extraction are described first of all to create the feature vector:

In order to allow sufficiently accurate detections, the preprocessing must map the Characteristic Scripture extensively, the resulting amount of data must be restricted in order to allow decoding, i.e. recognition, in an acceptable time (online and offline).

To describe handwriting sufficiently accurate in vectors it here needs editing of 11 features, which are described below. First, nine features are explained briefly, then describes the additional features are needed to better recognize cursive.

In this case, geometric features are examined the subject sliding window will be described briefly.

Sliding-Window: A systematic subdivision of the lettering after the pitch- and orientation-correction. These corrections are only useful when a handwriting-recognition is given. In a realization of a writer-recognition, this would be counter-productive, because writer-specific features would be eliminated.

It is pushed from left to right, a window of the text line. The height of the window is based on the font height, the window width normally is 4 pixels that overlap two pixels. These parameters are adjustable. It is shown that an increase of 14 pixels width and 13 pixels overlap window to significantly better results.

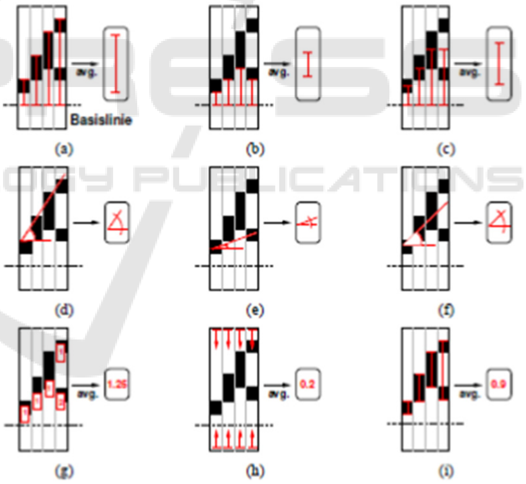


Figure 2: Previous features (Wienecke, M., Fink, G. A., and Sagerer, G., 2005).

(a) Mean distance of the upper outline of the type to the baseline

(b) Mean distance of the lower outline of the type to the baseline

(c) Average distance between the y-coordinates of the focal points columns

(d) Orientation of the upper contour

(e) Orientation of the lower contour

(f) Orientation of the course of columns priorities

(g) Average number of vertical header background transitions

(h) Mean Number of font pixels per image column

(i) Average number of font pixels between the upper and lower outline of the type

In the used system 10 feature-extractions are realized actually. The 10th feature is merely a small expansion, therefore it is not mentioned here.

As is evident from the representation of the feature extraction described, the position (s) and the angle of the lines (the) possibly multiple passage is not taken into account.

2 NEW FEATURE STRATEGIES

Therefore, two new features have to be introduced, which can occur more than once.

1. Passage distance from the base line (k)
2. Orientation of the passage line for the parallel to the base line (j)

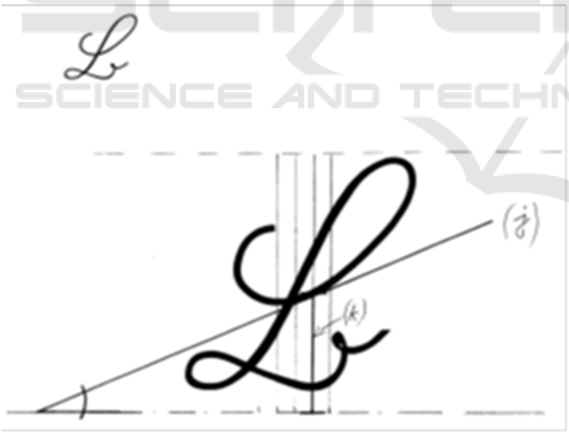


Figure 3: Principle of a new feature.

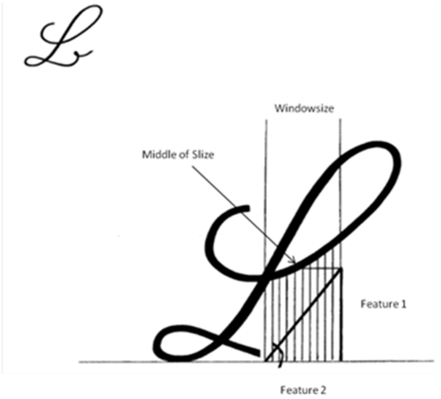


Figure 4: The example with a puncture in the realization.

The feature 1 (k) and the feature 2 (j) have been implemented and the results are evaluated.

Within the framework of the tests it was stated that people must deviate from the conventional procedure of the preprocessing. Usually a correction of the writing level (Slope), the letter inclination (Slant) as well as the standardization of the font size (Scale) is carried out in the case of the word recognition. Slope as well as Scale are used wider, Slant not.

Reason: The intersection characteristics are falsified by the erection of the letters in such a way that the recognition result is worsened.

Other considerations:

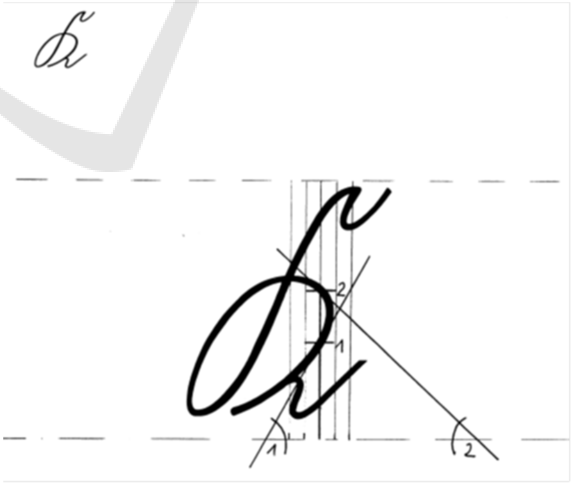


Figure 5: Multiple middle passages are possible, see German cursive capital letter K.

The realization of the extended intersection number takes place in a later publication.

Further feature:

Furthermore, it is considered whether the relationship between number of pixels of the top and bottom edge of the font is set in each pixel gap in relation to better accommodate to average values of the whole window, the baseline differences in the analysis. The further consideration and the implementation will be published in another publication.

Description of the recognition system

As a first step, a vector quantization is performed. The object is to reduce the data amount. For this reason, vectors from the input space are mapped to typical representatives, so similar data vectors are grouped together. The goal is to determine via cluster analysis accumulation ranges of unknown data distributions and to describe. Are formed mean vectors that are then representative of a statistical quantization.

Set Y = y1, y2, ... yN prototype vectors yi is also referred to as a codebook.

A typical quantization algorithm is the K-means algorithm today.

A brief description:

1. Initialization: Random selection of k cluster centers or the first N vectors of the sample.
2. Assignment: Each object is assigned to him closest cluster center, i.e. determination of optimal reproduction vector in the current codebook.
3. Recalculation: It will recalculate the cluster centers for each cluster, i.e. determination of a new code book.
4. Repeat: If the assignment changes, go to step 2, otherwise finished.

Scripture modeling

The static signature modeling is implemented via semi-continuous hidden Markov models. Under modeling is understood to mean the training of the training material. There models are so trained that correspond to the training material. Under decoding is understood to mean the detection, so the determination of the probability that the test pattern corresponding to a model. When modeling with HMM (Hidden Markov Model), we consider two modeling components:

The Writing Model: Hidden Markov Model

On word or letter level:

Language model (lexicon)

Markov Chain Model (n-gram model (Brakensiek, A., Rottland, J., and Rigoll, G., 2002))

The combination of both models provides a powerful system for the representation of handwriting. The parameters of the models can be predicted automatically.

Decoding (recognition):

So-called Decoding the combined model, meaning the optimal path through the combined state space. It is achieved optimal segmentation and classification in a continuous system. The Markov model concept describes the analysis of sequential data: The HMM as a statistical model is now "state of the art".

The recognition model:

Sequence of symbols (such as words)

w: Implementation in sequence of feature vectors X.

Objective of the recognition process:

Aim of the recognition process: Find the sequence which maximizes the posterior-probability P(w|X) of the symbol sequence of the given data.

Bayestheorem

$$\hat{w} = \arg\max_{w} P(w|X) = \arg\max_{w} \boxed{\frac{P(w)P(X|w)}{P(X)}}$$

$$= \arg\max_{w} P(w)P(x|w)$$

Application Bayestheorem:

posterior probability P (w | x) is implemented in the form in which the two component model of a Markov model are obvious:

P (w) = language model probability of symbol sequence w
n-gram model (Language Model).
P (w | X) = probability of observing the symbol sequences as features X, according to the writing model, namely the Hidden Markov Model (HMM).

Modeling:
HIDDEN MARKOV MODEL

General:
Using a hidden Markov model (Vinciarelli, A., Bengio, S., and Bunke, H., 2004) is trying to determine the probability with which a given feature

sequence (simply put: one-word commands) belongs to a particular word. For each word in the lexicon, an HMM is (i.e. for each word to be recognized) created. Once an input value (word) is present (observation), this is quasi compared with each HMM. The best result represents the recognized word. Verification with HMM is a statistical method. Determining the degree of match between the test sample and the reference as a probability.

2-piece stochastic process:

First Part: State machine: Describes probabilities of the transitions of the states.

Second Part: modeling the output of the font pattern when entering the respective state output probability distributions.
Allocation of the state to the magazine segment.

Feature of the chain is determined by the characteristics of the segments.

Training: Baum-Welch algorithm (optimization of HMM)

Improvement of a given model in response to certain sample data (training pattern) in such a way that the optimized model generates the training set with equal or greater probability.

Decoding: Viterbi algorithm

Calculation of the optimal path through the state sequence (Viterbi path). So the maximum probability of generating the observation sequence.

Is used to "discharging" of the hidden state sequence, that generates a maximum likelihood, a valid sequence of outputs, is given by the model.

3 RESULTS

Description of the database:

To make recognition systems comparable, this system must come on a standardized database to the application. In the case of the Old German handwriting (Kurrent) this is not possible, since such a database does not exist. It had to be redesigned a database. This database consists of approximately 6000 samples. This is made up of collections of 6 writers, here also some of Alexander von Humboldt. Characteristic lines and characteristic features of the writing thickness allow conclusions on the writer.

First, the test was carried out with 110 test images. The recognition rate was 0.4234. In order to present an apparent delta, more 440 test images were deducted from the training data. Then the model 11 and the model features 9 characteristics was examined. The window setting was w8X4 .

The term dimension in this context means: features or characteristics.

Table 1: Without correction (Slope, Slant, Scale).

Testimages	10-Dimensions	12-Dimensions
110	0.4234	0.4545
550	0.327	0.425

In the model with two additional features (12 dimensions), the difference is clear. At 550 test images, the recognition rate was increased by approximately 0.1!

Table 2: Without correction (Slant).

Testimages	10-Dimensions	12-Dimensions
110	0.5834	0.6038
550	0.5195	0.6284

Table 3: With correction (Slope, Slant, Scale).

Testimages	10-Dimensions	12-Dimensions
110	0.5636	0.554545
550	0.549091	0.516364
1100	0.534545	0.495455

4 CONCLUSIONS

Other Windows settings are investigated.

(e.g. w14X13).

In order to improve the recognition result, the number of samples must be increased noticeably. In this publication should be shown merely, that through an extension of the feature extractions, here specific for Old German handwriting (Kurrent), an improvement of the recognition is reached.

REFERENCES

Fink, G. A., 2008. *Markov Models for Pattern Recognition, From Theory to Applications.* Springer, Heidelberg.

Wienecke, M., Fink, G. A., and Sagerer, G., 2005. *Toward Automatic Video-based Whiteboard Reading. Int. Journal on Document Analysis and Recognition,* vol.

7(2–3):188–200.

Steinke, K.-H.; Dzido, R.; Gehrke, M.; Prätel, K., 2008. *Featurerecognition for herbarium specimens (Herbar-Digital),* Proceedings of TDWG, Perth.

Vinciarelli, A., Bengio, S., and Bunke, H., 2004. *Offline Recognition of Unconstrained Handwritten Texts using HMMs and Statistical Language Models. IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26(6):709–720.

Brakensiek, A., Rottland, J., and Rigoll, G., 2002. *Handwritten Address Recognition with Open Vocabulary Using Character N-Grams.* In *Proc. Int. Workshop on Frontiers in Handwriting Recognition*, pages 357–362. Niagara on the Lake, Canada.

Fink, G. A., 1999. *Developing HMM-based Recognizers with ESMERALDA*. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, eds., *Text, Speech and Dialogue*, vol. 1692 of *Lecture Notes in Artificial Intelligence*, pages 229–234. Springer, Berlin Heidelberg.