GMM-based Handwriting Style Identification System for Historical Documents

Fouad Slimane[1,3], Torsten Schaßan[2], Volker Märgner[1]
[1]*Institute for Communications Technology (IfN), Technische Universität Braunschweig*
[2]*Herzog August Bibliothek Wolfenbüttel (HAB)*
Braunschweig, Germany
[3]*School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne*
Lausanne, Switzerland
Email: {slimane,maergner}@ifn.ing.tu-bs.de, schassan@hab.de

***Abstract*—In this paper, we describe a novel method for handwriting style identification. A handwriting style can be common to one or several writer. It can represent also a handwriting style used in a period of the history or for specific document. Our method is based on Gaussian Mixture Models (GMMs) using different kind of features computed using a combined fixed-length horizontal and vertical sliding window moving over a document page. For each writing style a GMM is built and trained using page images. At the recognition phase, the system returns log-likelihood scores. The GMM model with the highest score is selected. Experiments using page images from historical German document collection demonstrate good performance results. The identification rate of the GMM-based system developed with six historical handwriting style is 100%.**

***Keywords*-handwriting style; GMMs; local features; sliding window; historical German document collection**

I. Introduction

Historical document collections present a difficult challenges for information retrieval. The absence of consistent orthographic conventions in historical document reveals many handwriting styles and make the text analysis and recognition difficult. Most of the time, for each type of document we need a specific method for information retrieval.

Historical handwriting styles are a carefully designed, efficient way of forming letters and numbers. Each style has its own character or fits certain needs. Styles change over time and every days new ones emerge. Sometimes, the styles used for writing legal and administrative documents were slightly different from formal book hands used for literary works, but shared many characteristics. Figure 1 shows some example pages from the used dataset. This figures (a), (b), (c), (d), (e) and (f) show an example of page written respectively with the *Carolingian Minuscule* style, *Textualis* style, *Bastarda* style, *Cursiva recentior / jüngere gotische Kursive* style, *Insulare Minuscule* style, *Carolino Gothica* style. All details about each style will be presented in Section II.

The objective of the handwriting style identification problem is to automatically determine the identity of the style of a handwritten document from among a set of possible historical styles. Many historical books were written by different writers but with the same style. The writer identification in this case is a difficult task. It is better to recognize the style and then the writer. This is why, in this paper we propose to identify the handwriting style. The style identification could help the handwritten recognition systems to draw conclusions about the shape of the characters given facts about the writers and could also be used for purposes of writer identification, and the clustering of books written with the same handwriting style.

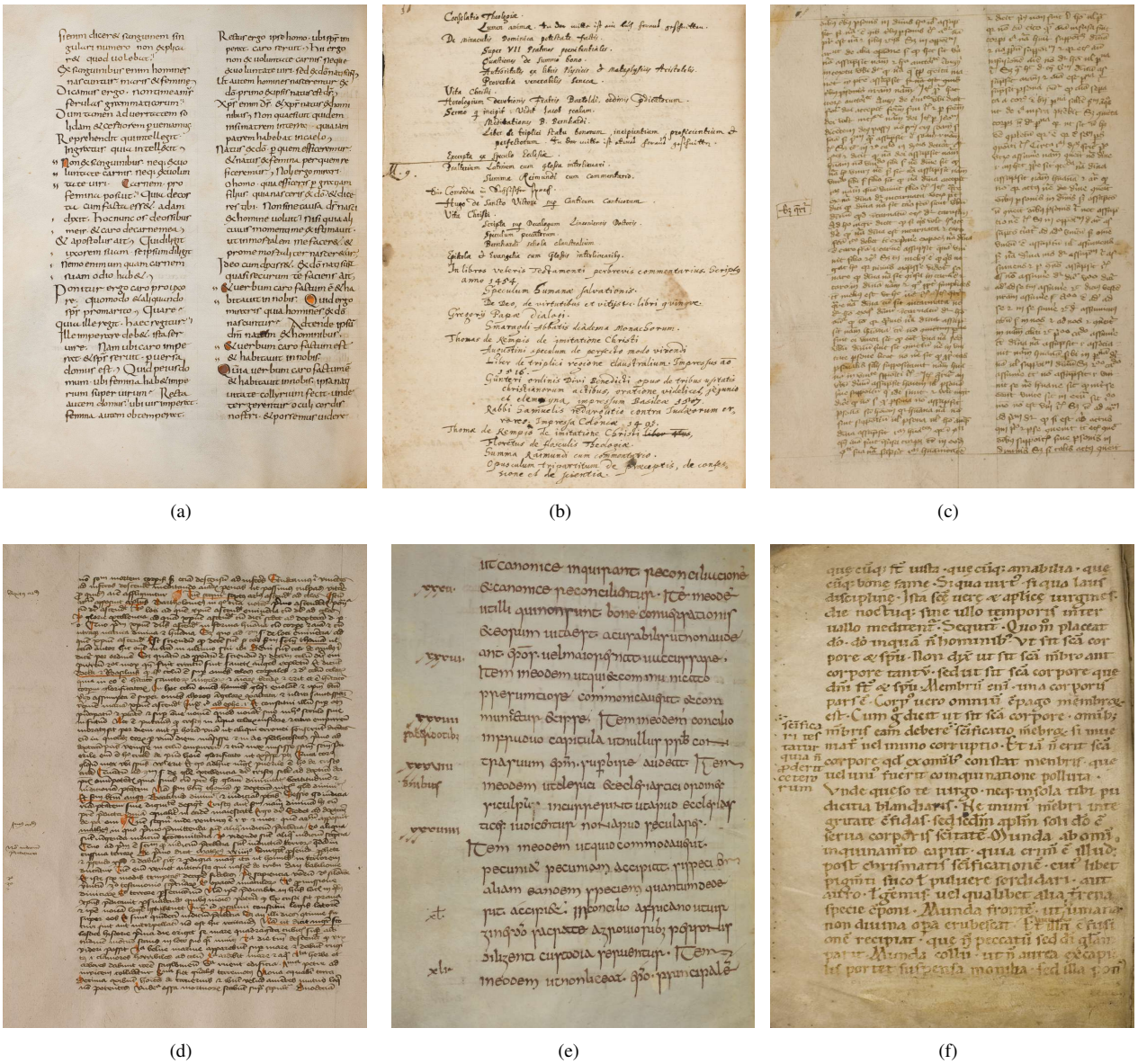
Handwriting is usually learned by copying a formal system. Although changing somewhat over time, ones handwriting style typically originates from a particular historical style. The problem of categorizing handwriting styles has not been considered widely. Only some works are published on this field. Some approaches used neural network or hierarchical clustering techniques to find a set of handwriting families or clusters [1], [2], [3]. Cluster is made to represent one prototypical allograph. The allograph extraction methods have been shown to improve the performance of recognizers. The basis of the allograph categorization is the concept of the stroke. The segmentation into strokes is made in different ways (see [4]).

Crettez [5] presents a system for handwriting style clustering. Considering the fact that writers draw their strokes in some directions more frequently than in others, which decides the general slant and spread of the handwriting, the author finds that each writer generally has four "preferential directions". The directions are used to create a number of handwriting categories that can be used for adapting recognizers.

All published works have some limitation and do not cover historical documents and their specific characteristics. One of the novelties in this paper is to present robust handwriting style system for historical document, to not segment page images into blocks and lines, and to use a simple approach based on Gaussian Mixture Models (GMMs).

The rest of this paper is organized as follows. Section II describes the corpus used for our experiments. Section III details the handwriting style identification system. Section IV is dedicated to the experimental results and it is

978-1-4799-5934-1/14/$31.00 ©2014 IEEE



(a) (b) (c)

(d) (e) (f)

Figure 1. Example pages of our corpus: (a) Carolingian Minuscule - page from the "Augustinus: In Iohannis evangelium tractatus 123" book (9 century), (b) Textualis - page from the "Bibliothekskataloge aus Helmstedt und Wolfenbttel" data (17 century), (c) Bastarda - page from the "Quaestiones in libros III et IV sententiarum"' book (15 century), (d) Cursiva Recentior - page from the "Floretus cum commento" data set (15 century), (e) Insulare Minuscule - page from the "Admonitio generalis. Sermones de symbolo. Expositiones orationis dominicae" data set (8/9 century) and (f) Carolino Gothica - pages from the "Institutio sanctimonialium Aquisgranensis. Ps.-Eusebius Gallicanus" data set (12 century).

followed by a conclusion and future work.

II. Corpus: Historical German Document Collection

As part of a collaboration with the Herzog August Bibliothek Wolfenbüttel (HAB)[1] in Germany, we tested our system with scanned page images from books written between the 8^{th} and the 17^{th} centuries with different writers and different handwriting styles. The used dataset is available under the HAB website. Six different historical handwriting styles are used in this work:

1) **Insulare Minuscule**: text written in Insulare minuscule commonly use large initial letters surrounded by red ink dots. Letters following a large initial at the start of a paragraph or section often gradually diminish in size as they are written across a line or a page, until the normal size is reached, which is called a "diminuendo" effect, and is a distinctive

[1] http://www.hab.de/

insular innovation, which later influenced continental illumination style. Letters with ascenders (*b*, *d*, *h*, *l*, etc.) are written with triangular or wedge-shaped tops. The bows of letters such as *b*, *d*, *p*, and *q* are very wide. The script uses many ligatures and has many unique scribal abbreviations, along with many borrowings from Tironian notes. The handwriting documents used in this work are from the 8^{th} century and available in this link[2].

2) **Carolingian Minuscule**: Carolingian minuscule was uniform, with rounded shapes in clearly distinguishable glyphs, disciplined and above all, legible. Clear capital letters and spaces between words became standard in Carolingian minuscule, which was one result of a campaign to achieve a culturally unifying standardization across the Carolingian Empire. Carolingian script generally has fewer ligatures than other contemporary scripts. The early period of the script, during Charlemagne's reign in the late 8^{th} and early 9^{th} century, still has widely varying letter forms in different regions. The script flourished during the 9^{th} century, when regional hands developed into an international standard, with less variation of letter forms. The script began to decline slowly after the 9^{th} century. In the 10^{th} and 11^{th} centuries, ligatures were rare, and ascenders began to slant to the right and were finished with a fork. By the 12^{th} century, Carolingian letters became more angular and were written closer together, less legibly than in previous centuries. The handwriting documents used in this work are from the 9^{th} century and available in those links[3] [4] [5].
3) **Carolino Gothica**: The handwriting documents used in this work are from the 12^{th} century and available in this link[6].
4) **Bastarda**: Bastarda (or bastard or lettre *bâtarde* in French) was a Gothic script used in Germany and France during the 14^{th} and 15^{th} centuries. These scripts were used to provide a simplified letter that was appropriate for the copying of documents of minor value or importance. The handwriting documents used in this work are from the 15^{th} century and available in this link[7].
5) **Cursiva Recentior / jüngere gotische Kursive**: Cursiva recentior was by far the most widely used type of script in the 14^{th} and the 15^{th} centuries. It is found in many manuscripts all over Europe, from the most informal ones (these are the majority) to the various calligraphic versions which developed in different countries, and the large group of codices in Cursiva Libraria. The handwriting documents used in this work are from the 15^{th} century and available in those links [8] [9].
6) **Textualis**: Textualis, also known as textura or Gothic bookhand, was the most calligraphic form of black letter, and today is the form most associated with "Gothic". Johannes Gutenberg carved a textualis typeface including a large number of ligatures and common abbreviations, when he printed his 42-line bible. However, the textualis was rarely used for typefaces afterwards. Some characteristics of this style are:
   - tall, narrow letters, as compared to their Carolingian counterparts.
   - letters formed by sharp, straight, angular lines, unlike the typically round Carolingian; as a result, there is a high degree of "breaking", i.e. lines that do not necessarily connect with each other, especially in curved letters.
   - a related characteristic is the half *r*, the shape of *r* when attached to other letters with bows; only the bow and tail were written, connected to the bow of the previous letter. In other scripts, this only occurred in a ligature with the letter *o*.
   - similarly related is the form of the letter d when followed by a letter with a bow; its ascender is then curved to the left, like the uncial d. Otherwise the ascender is vertical.

   The handwriting documents used in this work are from the 17^{th} century and available in this link[10].

Figure 1 illustrates one sample page of each of these handwriting styles.

III. System Description

As illustrated in Figure 2, the proposed system includes two parts. The first part is a front-end for the preprocessing of the images and for the feature extraction. As page images in the used database are scanned in color, they are transformed to gray levels using an adapted threshold to each page image.

The second part computes likelihood estimators of each handwriting style model. For each handwriting style, a GMM is built and trained with data coming from that handwriting style only. To train the GMM, a set of features are extracted from a page image using a combined horizontal and vertical sliding window. The sliding window has the size of 50×50 pixels, moves from left to right, top to down, horizontally with a shift of 10 pixels and vertically with a shift of 20 pixels. The horizontal and vertical window widths

[2]http://diglib.hab.de/mss/496a-helmst/start.htm
[3]http://diglib.hab.de/mss/10-weiss/start.htm
[4]http://diglib.hab.de/mss/14-weiss/start.htm
[5]http://diglib.hab.de/mss/18-weiss/start.htm
[6]http://diglib.hab.de/mss/877-helmst/start.htm
[7]http://diglib.hab.de/mss/278-helmst/start.htm
[8]http://diglib.hab.de/mss/384-helmst/start.htm
[9]http://diglib.hab.de/mss/82-4-aug-2f/start.htm
[10]http://diglib.hab.de/mss/27-2-aug-2f/start.htm

of the sliding window and the shift pixels were optimized in an independent validation experiment.

GMM models are trained using the Expectation Maximization (EM) algorithm [6]. The recognition is performed through a score comparison of the trained Gaussian mixture models. The computational complexity of our approach increases linearly with the number of handwriting styles to be identified.

A. Preprocessing and Feature Extraction

In the sliding window, seventeen features are extracted. Our choice of features is based on several experiments and using an incremental test selection. In the following the used features are presented:

1) Density of pixels in the window
2) Vertical position of the gravity center in the whole window (W). The result is normalized by the height h of the window as presented in Equation 1.

$$f = \frac{G_y(W)}{h} \tag{1}$$

3) Mean of 12 Zernike moments computed from the window [7]. To evaluate Zernike moments, the image (or region of interest) is first mapped to a unit circle using polar coordinates, where the center of the image is the origin of the unit circle. Pixels falling outside the unit circle are not taken into consideration. They are also robust to noise and grey level variations of shapes like anti-aliasing artifacts. Zernike introduced a complete orthogonal set $\{V_{nm}(x,y)\}$ of complex polynomials over the polar coordinate space inside a unit circle (i.e., $x^2+y^2=1$) as following:

$$V_{nm}(x,y) = V_{nm}(\rho,\theta) = R_{nm}(\rho)\exp^{jm\theta} \tag{2}$$

where $j=\sqrt{-1}$, $n \geq 0$, m is an integer, $|m| \leq n$ and $n-|m|$ is even, ρ is the shortest distance from the origin to the pixel (x,y), θ is the angle between the vector ρ and the x-axis in counter-clockwise direction, $R_{nm}(\rho)$ is the orthogonal radial polynomial defined by:

$$R_{nm}(\rho) = \sum_{s=0}^{(n-|m|)/2} (-1)^s \frac{(n-s)!}{s!(\frac{n+|m|}{2}-s)!(\frac{(n-|m|}{2}-s)!}\rho^{n-2s} \tag{3}$$

In the case of digital image I, the Zernike moment of order n and repetition m is defined as:

$$A_{nm} = \frac{n+1}{\pi}\sum_x\sum_y I(x,y)V^*_{nm}(\rho,\theta), \tag{4}$$

where * presents the complex conjugate.

4) Mean of vertical projection normalized by the window width
5) Mean of horizontal projection
6) Standard deviation of vertical projection normalized by the window width
7) Standard deviation of horizontal projection
8) Derivate of vertical projection vector profile normalized by the window width

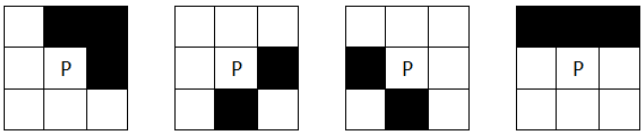


Figure 3. The used typological masks around a background pixel P

9) Derivate of horizontal projection vector profile
10) Mean of vertical runs
11) Mean of horizontal runs
12) Standard deviation of vertical runs
13) Standard deviation of horizontal runs
14) - 17) Number of white pixels according to each of the four typological masks shown in Figure 3 in the window normalized by the size of the window (4 features)

Using the combined horizontal and vertical sliding window technique, no segmentation into paragraph line and letters is made and the page image is transformed into a sequence of feature vectors. The number of rows corresponds to the number of components of each feature vector, and the number of columns corresponds to the number of analysis windows.

The sequence of seventeen-dimensional feature vectors thus obtained from each page image is used to train the GMMs. As a result of the training procedure, we obtain for each handwriting style a GMM that is specially adapted to this specific style.

B. Gaussian Mixture Models for Handwriting Style Identification Modeling

Gaussian Mixture Models (GMMs) were used in many domains such as font recognition [8], [9], label image verification [10], baseline estimation [11], writer identification [12], etc. In this work GMMs are used to model the handwriting style of each historical book(s). We modeled the distribution of the feature vectors using Gaussian mixture density.

For a D-dimensional feature vector x the mixture density for each writer is defined as:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{5}$$

The density is a weighted linear combination of M uniform Gaussian densities $p_i(x)$, each parametrized by a $D \times 1$ mean vector μ_i and a $D \times D$ covariance matrix $\sum_i$. The parameters of a handwriting style's density model are presented as $\lambda = \{w_i, \mu_i, \sum_i\}$ where the mixture weights w_i sum up to one. To simplify the computation, we make the hypothesis that the feature vectors coefficients are not correlated. The covariance matrix is then simplified to a diagonal matrix. This approximation is classically done when using GMM and have shown that diagonal matrix perform better than full covariance matrices [13].

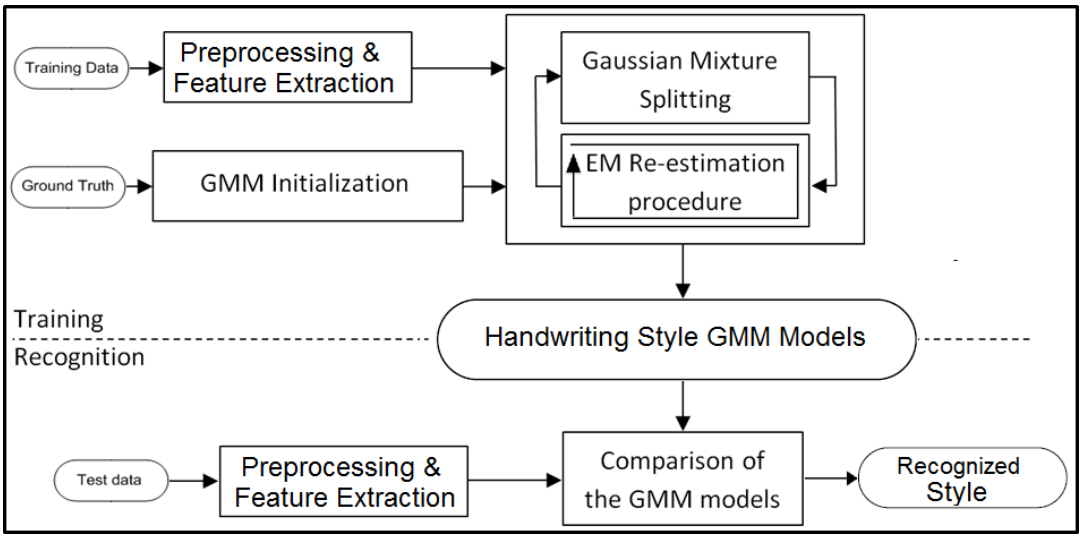


Figure 2. GMM based handwriting style identification system

During training, the iterative EM algorithm is used to refine the GMM parameters (component weights, means and variances) to increase the likelihood of the estimated model for the observed feature vectors [8]. In our experiments, we used the EM algorithm to build the models by applying a binary splitting procedure to increase the number of Gaussian mixtures through the training procedure. As our objective is here to maximize the recognition performance, we have chosen to use 512 Gaussians as reference for our handwriting style identification system. In the next section, we present results using different number of Gaussians.

Considering the hypothesis of feature vector independence, the log-likelihood of a model λ for a sequence of feature vectors, $X = \{x_1, \ldots, x_N\}$ is computed as follows:

$$log\ p(X|\lambda) = \sum_{i=1}^{N} log\ p(x_i|\lambda) \quad (6)$$

where $p(x_i|\lambda)$ is computed according to Eq. 5.

During decoding, a page to be classified is presented to the GMM of each handwriting style. Each GMM outputs the log-likelihood score and the standard deviation for the given page image. The log-likelihood scores are sorted in decreasing order and the page is assigned to the best ranked handwriting style.

System performances are evaluated in terms of handwriting style identification rates using an unseen set of page images.

Our GMM-based system is implemented using the Hidden Markov Model Toolkit (known as HTK Toolkit)[11] [14].

[11]http://htk.eng.cam.ac.uk/

IV. Handwriting style Identification Results and Discussion

To evaluate our system it is essential to collect handwritten documents whose style is known. The evaluation of our handwriting style identification system has been conducted using a historical German document collection with 6 styles. In all tests, identification rates have been evaluated at page level.

The experiments are based on page images. For each handwriting style we use about 35 pages for training and 15 for testing.

In a first set of handwriting style identification experiments, we tried to find the optimal number of Gaussians for the handwriting style models. We tested our system with 1, 2, 4, 8, ... , 512, 1024 Gaussians in the mixture, doubling the number of Gaussians after each 10 iterations of the EM algorithm. Figure 4 illustrates the evolution of the handwriting style identification rates as a function of the number of Gaussians in the models. The highest writer identification rate of 100% is achieved using 512 Gaussians. The handwriting style identification rate increase as the number of mixtures increases although using too many mixtures (>512) causes degradation of the performance due to over-training.

The results are very promising and can help for example to identify the writer of a questioned document or to improve text recognition. The global text-independent handwriting style identification rate of the GMMs-based system is 100%.

Due to the lack of available systems for comparison, we develop and compare the presented system with a k-Nearest-Neighbor (k-NN) based system with k=1 using the same features. To identify the handwriting style of a page, we compute the euclidean distance between this page and each page used in the training set. The resulting style is equal to

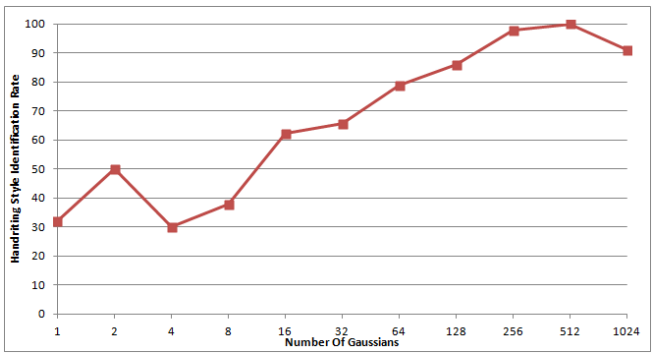


Figure 4. Handwriting style identification rate as a function of the number of Gaussians in the GMM-style model.

the class of the nearest page contained in the training set. The identification rate with the K-NN based system is 93.51%. This system is much faster but performs worse results than the GMM-based one.

V. Conclusion

In this paper we use Gaussian mixture models to address the task of text-independent handwriting style identification applied to historical document. We model six handwriting style distribution using GMMs and only seventeen local features computed from a combined horizontal and vertical sliding window. When presented with a page of unknown origin, each GMM outputs the log-likelihood score and standard deviation for the given input. We rank the log-likelihood scores of each model and choose the highest ranked handwriting style. A historical German document collection was used in the analysis and experiments. The accuracy results for different styles were presented and analyzed. The handwriting style identification rate of the GMMs-based system developed for six styles is 100%.

In the future, we will explore more features, optimize the used set, evaluate the system with more handwriting styles, and test other classifiers like Support Vector Machines (SVM) and Neuronal Networks like MLP for handwriting style identification.

Acknowledgment

This work has been supported by the Swiss National Science Foundation fellowship project PBFRP2_145898.

References

[1] L. Schomaker, G. Abbink, and S. Selen, "Writer and writing-style classification in the recognition of online handwriting," *IEE European Workshop on Handwriting Analysis and Recognition: A European Perspective*, pp. 1/1–1/4, Jul 1994.

[2] L. Vuurpijl and L. Schomaker, "Finding structure in diversity: a hierarchical clustering method for the categorization of allographs in handwriting," *International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, pp. 387–393, 1997.

[3] M. L. Bote-lorenzo, Y. A. Dimitriadis, and E. Gmez-snchez, "Allograph extraction of isolated handwritten characters," *Proc. of the Tenth Biennial Conference of the International Graphonomics Society (IGS)*, pp. 191–196, 2001.

[4] R. Plamondon and S. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, Jan 2000.

[5] J.-P. Crettez, "A set of handwriting families: style recognition," *Third International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, pp. 489–494 vol.1, Aug 1995.

[6] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," in *Royal Statistical Society Series B Methodological*, 1977, vol. 39, no. 1, pp. 1–38.

[7] F. Zernike, "Beugungstheorie des schneidenverfahrens und seiner verbesserten form, derphasenkontrastmethode (diffraction theory of the cut procedure and its improved form, the phase contrast method)," *In Physica*, vol. 1, pp. 689–704, 1934.

[8] F. Slimane, S. Kanoun, J. Hennebert, A. M. Alimi, and R. Ingold, "A study on font-family and font-size recognition applied to arabic word images at ultra-low resolution," *Pattern Recognition Letters*, vol. 34, no. 2, pp. 209 – 218, 2013.

[9] F. Slimane, S. Kanoun, A. M. Alimi, R. Ingold, and J. Hennebert, "Gaussian mixture models for arabic font recognition," *International Conference on Pattern Recognition (ICPR)*, pp. 2174–2177, 2010.

[10] M. Baechler, J.-L. Bloechle, and J. Hennebert, "Labeled images verification using gaussian mixture models," in *Proceedings of the 2009 ACM Symposium on Applied Computing*, ser. SAC '09. New York, NY, USA: ACM, 2009, pp. 1331–1335. [Online]. Available: http://doi.acm.org/10.1145/1529282.1529581

[11] F. Slimane, S. Kanoun, J. Hennebert, R. Ingold, and A. M. Alimi, "A New Baseline Estimation Method Applied to Arabic Word Recognition," *10th IAPR International Workshop on Document Analysis Systems (DAS 2012)*, 2012.

[12] F. Slimane and V. Maergner, "A New Text-Independent GMM Writer Identification System Applied to Arabic Handwriting," *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, p. to be published, 2014.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 13, pp. 19 – 41, 2000.

[14] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.