Pattern Anal Applic (2011) 14:87–92
DOI 10.1007/s10044-010-0178-6

SHORT PAPER

Automatic gender detection using on-line and off-line information

Marcus Liwicki · Andreas Schlapbach ·
Horst Bunke

Received: 10 February 2009 / Accepted: 25 February 2010 / Published online: 22 April 2010

Abstract In this paper, the problem of classifying handwritten data with respect to gender is addressed. A classification method based on Gaussian Mixture Models is applied to distinguish between male and female handwriting. Two sets of features using on-line and off-line information have been used for the classification. Furthermore, we combined both feature sets and investigated several combination strategies. In our experiments, the on-line features produced a higher classification rate than the off-line features. However, the best results were obtained with the combination. The final gender detection rate on the test set is 67.57%, which is significantly higher than the performance of the on-line and off-line system with about 64.25 and 55.39%, respectively. The combined system also shows an improved performance over human-based classification. To the best of the authors' knowledge, the system presented in this paper is the first completely automatic gender detection system which works on on-line data. Furthermore, the combination of on-line and off-line features for gender detection is investigated for the first time in the literature.

M. Liwicki
Knowledge Management Department, German Research Center
for AI (DFKI GmbH), Kaiserslautern, Germany

M. Liwicki (✉) · A. Schlapbach · H. Bunke
Institut für Informatik und angewandte Mathematik,
Universität Bern, Neubrückstrasse 10, 3012 Bern, Switzerland
e-mail: Marcus.Liwicki@dfki.de; liwicki@iam.unibe.ch

A. Schlapbach
e-mail: schlpbch@iam.unibe.ch

H. Bunke
e-mail: bunke@iam.unibe.ch

Keywords Gender detection · Handwriting analysis · Gaussian Mixture Models · Multiple classifier combination

1 Introduction

A population of individuals can often be partitioned into sub-categories based on various criteria. Dividing a population into sub-categories is interesting for numerous reasons, for example, if a researcher is only interested in one specific sub-category, or if specifically processing each sub-category leads to improved results. For example, in the field of face recognition, much research has been conducted on classifying a face image according to gender [24, 25] or to divide a population into individuals that wear glasses and individuals that do not [12]. Classification results up to 94% have been reported for such two-class problems.

For handwritten data there exist several possible criteria to define sub-categories. Whereas in KANSEI the sub-categories are based on feelings, emotions, and character traits [8], handwriting can also be divided into writer-specific sub-categories including gender, handedness, age and ethnicity [20]. Correlations between these sub-categories and handwriting features have been presented in [10]. Special interest has been focused on determining the gender of the writer. In [7], humans were asked to classify the writer's gender of a given handwriting sample. A classification rate of about 68% has been reported. Further studies in [2], which inlcude a detailed analysis of the rater's background, reported results in the same range.

Especially the classification of gender from handwriting has been a research topic for many decades [3, 17, 23]. However, there exist conflicting results ranging from slightly more than 50% to more than 90%. An overview of several manual approaches detecting gender from handwriting can be found in [9]. This thesis tries to semi-automatically classify the handwriting (while it is done automatically in this paper).

Little work exists on automatically identifying sub-categories, such as gender or handedness, from handwriting. In [4], a system for classifying the handwriting based on images of individual letters is presented. Results of 70.2% for gender classification and 59.5% for handedness have been achieved. If longer texts are available and multiple classifier approaches are applied even better results are reported [1]. However, these systems are restricted to the off-line case and either the transcription of the text has to be known or even identical texts have to be provided by all writers.

In this paper, we present a system that classifies gender of on-line, Roman handwriting. This problem is a two class problem, i.e., male/female. On-line handwriting means that temporal information about the handwriting is available. As the handwriting is unconstrained, any text can be used for classification. Two sets of features are investigated in this paper. While the first feature set is based on on-line data, the second set of features is extracted from off-line images generated from the on-line data. We applied Gaussian Mixture Models (GMMs) to model the classes. In our experiments, the classifier working with the on-line features outperforms the off-line classifier. Furthermore, we combined both feature sets and investigated several combination strategies. Using the maximum rule in the combination turned out to give the best results. The final gender detection rate on the test set is 67.57%, which is significantly higher than the performance of the on-line and off-line system with about 64.25 and 55.39%, respectively. To the best of the authors' knowledge, the system presented in this paper is the first completely automatic gender detection system which works on on-line data. Furthermore, the combination of on-line and off-line features for gender detection is investigated for the first time in the literature. For the purpose of comparison, also an experiment with humans classifying the same on-line data set is performed.

2 Data acquisition and feature extraction

To acquire the handwritten data, the eBeam[1] interface is used. It generates a sequence of (x, y)-coordinates representing the location of the tip of the pen together with a time stamp for each location. The frame rate of the recordings varies from 30 to 70 frames per second. An illustration of the recording process is shown in Fig. 1.

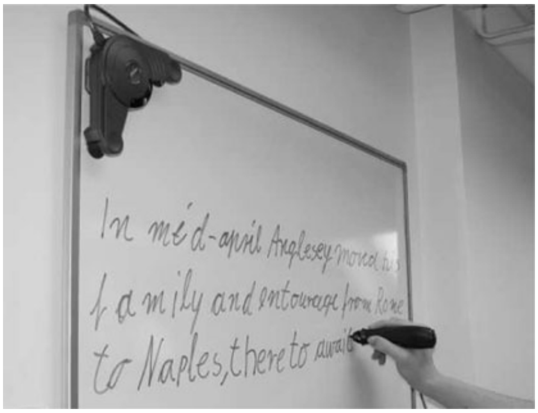

In mid-april Anglesey moved his
family and entourage from Rome
to Naples, there to awai

Fig. 1 Illustration of the recording process

The normalization and feature extraction is motivated from previous work in writer identification. Both, the on-line and off-line feature sets used in this paper have shown excellent performance on the writer identification and verification task [22, 21]. The gender identification task is related to the task of writer identification and can be posed as a two-class problem ("female" writer vs. "male" writer). Therefore it is reasonable to apply the normalization and feature extraction methods from [22] and [21] to the gender identification task.

For preprocessing, we divide each text line into sub-parts. For each sub-part the skew angle is corrected to horizontally align the text and a size normalization is performed [22].

The feature set used in this experiment contains on-line features as well as features extracted from an off-line representation of the on-line data. (The number in round brackets behind the name of a feature will indicate the number of individual features.) For a given stroke s consisting of points p_1 to p_n, the following 18 on-line features for each consecutive pair of points (p_i, p_{i+1}) are computed (see Fig. 2 for an illustration): speed (1); writing direction (2); curvature (2); normalized x- and y-coordinate (2); speed in x- and y-direction (2); overall acceleration (1); acceleration in x- and y-direction (2); log curvature radius (1), which is the length of the circle which best approximates the curvature at the point p_i [19], vicinity aspect (1), vicinity curliness (1); vicinity linearity (1); and vicinity slope (2) [11]. Furthermore, the following pseudo off-line features are computed using a two-dimensional matrix representing an off-line version of the data. The matrix is obtained by projecting the on-line strokes on the two-dimensional plane. ascenders/descenders (2), i.e., the number of points above/below the corpus line whose x-coordinates are in the vicinity and context map (9), i.e., the two-dimensional vicinity of the point is divided into three regions for each dimension. The number of black

[1] eBeam System by Luidia, Inc., http://www.e-Beam.com.

points in each region is taken as a feature value. Overall the feature set consists of 29 features [15].

The off-line system computes the features by applying a sliding window which moves in writing direction and extracting the following values at each window position: mean gray value of the pixels (1), center of gravity (1), second order vertical moment of the center of gravity (1), positions of the uppermost and lowermost black pixels (2), rate of change of these positions (2) (with respect to the neighbouring windows), number of black–white transitions(1) between the uppermost and lowermost pixels, proportion of black pixels(1) between the uppermost and lowermost pixels For a more detailed description of the off-line features, see [15].

3 Gaussian Mixture Models

We use Gaussian Mixture Models (GMMs) to model the handwriting of each sub-category of the underlying population. The distribution of the feature vectors extracted from a sub-category's handwriting is modeled by a Gaussian mixture density. For a D-dimensional feature vector $\mathbf{x}$ the mixture density for a specific sub-category is defined as

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

where the mixture weights w_i sum up to one. The mixture density is a weighted linear combination of M uni-modal Gaussian densities $p_i(\mathbf{x})$, each parametrized by a $D \times 1$ mean vector μ_i and a $D \times D$ covariance matrix C_i:

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|C_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_i)'(C_i)^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right\}. \tag{2}$$

The parameters of a sub-category's density model are denoted as $\lambda = \{w_i, \ \mu_i, \ C_i\}$ for all $i = 1, \ldots, M$, which completely describes the model.

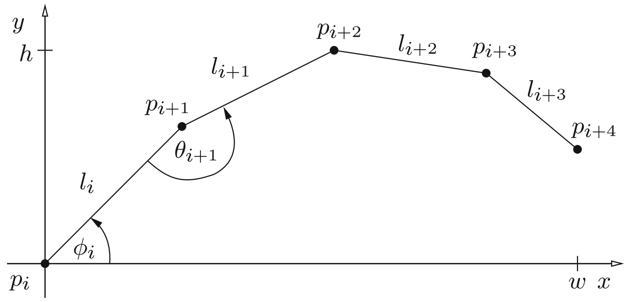


Fig. 2 Features extracted from the on-line handwriting

While the general model supports full covariance matrices, often only diagonal covariance matrices are used. An example of the two dimensional case is shown in Fig. 3. This simplification is motivated by the following observations: first, theoretically the density modeling of an M dimensional full covariance matrix can equally well be achieved using a larger order diagonal covariance matrix. Second, diagonal covariance matrices are computationally more efficient than full covariance matrices, and third, diagonal matrix GMMs outperformed full matrix GMMs in various experiments [18].

Instead of training a sub-category model from scratch for every sub-category, we obtain the models of the sub-categories from a Universal Background model (UBM). The basic idea is to derive the sub-category's model by updating the well-trained parameters from the UBM. In a first step, all data from all writers are used to train a single, writer independent UBM. Training is performed with the expectation–maximization (EM) algorithm [6] In the second step, for each sub-category a sub-category dependent model is build by updating the parameters in the UBM via adaptation using all training data from this sub-category. Therefore a modified version of the EM algorithm is used, which is based on the Maximum a Posteriori (MAP) principle.

For training, variance flooring is employed to avoid an overfitting of the variance parameter [16]. The idea of variance flooring is to impose a lower bound on the variance parameters as a variance estimated from only few data points can be very small and might not be representative of the underlying distribution of the data [16].

For adaptation the new statistical estimates are then combined with the old statistics from the UBM mixture parameters using a data-dependent mixture coefficient. This adaptation coefficient (called MAP adaptation factor) controls the adaptation process by emphasizing either on

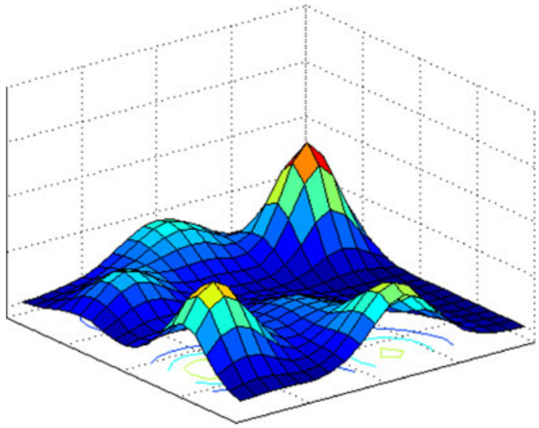


Fig. 3 A two-dimensional GMM consisting of a weighted sum of six uni-modal Gaussian densities

the well-trained data of the UBM or on the new data when estimating the parameters [18].

During decoding, the feature vectors $X = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$ extracted from a text line are assumed to be independent. The log-likelihood score of a model λ for a sequence of feature vectors X is defined as

$$\log p(X|\lambda) = \sum_{t=1}^{T} \log p(\mathbf{x}_t|\lambda), \tag{3}$$

where $p(\mathbf{x}_t|\lambda)$ is computed according to Eq. 1.

4 Combination

After decoding, each classifier returns a log-likelihood score, i.e., the on-line classifier returns $ll_{\text{on-line}}$ and the off-line classifier returns $ll_{\text{off-line}}$. Having on-line and off-line classification systems available, it may be beneficial to combine both systems. From such a combination, an improved performance can be expected. For a general overview and an introduction to the field of multiple classifier systems (MCS) see [13].

To combine the results of the on-line and the off-line classifier, the following standard rules for the classifier combination on the score level are applied [5]: *Average rule* The scores of both classifiers are averaged: $ll_{\text{sum}} = \frac{1}{2}(ll_{\text{on-line}} + ll_{\text{off-line}})$. *Maximum Rule* The largest score is chosen: $ll_{\max} = \max(ll_{\text{on-line}}, ll_{\text{off-line}})$. *Minimum Rule* The smallest score is chosen: $ll_{\min} = \min(ll_{\text{on-line}}, ll_{\text{off-line}})$. The range of log-likelihood scores of both classifiers vary greatly. Therefore, before combination the results of both classifiers are normalized in respect to mean and standard deviation. Due to the fact that only two classifiers are used, other combination rules such as the median rule or voting are not applicable in this case.

5 Experiments and results

The experiments have been conducted on the IAM-OnDB [14], a large on-line handwriting database acquired from a whiteboard.[2] This database consists of data from more than 200 writers with eight handwritten texts per writer. Each text consists of seven text lines on average. The classification task is to identify the correct gender for a given text line.

For the task of gender classification we randomly selected 40 male and 40 female writers for training the classifiers, 10 male and 10 female writers for the validation of meta-parameters, and 25 male and 25 female writers for testing the final system. This assures that both classes are equally distributed in all sets, the training, the validation, and the test set. Note that these sets were the same for all experiments described in the remainder of this section.

For the GMM the number of Gaussian mixture components G were optimized between 1 and 250. Next, the variance flooring factor φ was varied between 0.001 and 0.011 in steps of 0.002. Furthermore, the MAP adaptation factor α was varied from full adaptation (which corresponds to $\alpha = 0$ in the Torch library) to no adaptation ($\alpha = 1$) in steps of 0.2. All these optimization operations were carried out on the validation set.

The optimization of the MAP adaptation factor and the number of Gaussian mixtures is illustrated in Fig. 4. In this figure, the MAP adaptation factor and the number of Gaussian mixtures are plotted together with the corresponding classification rate obtained on the validation set. To highlight the differences we have added dashed lines indicting certain performance levels (54, ..., 64%). The maximum is reached at $\alpha = 0.8$ and $G = 220$. In order to improve the clarity of this figure, Table 1 shows the performance on the validation set near the local maximum. These results show that, first, a higher number of Gaussian mixture components leads to significantly higher classification results until a maximum of 220 Gaussian mixture components is reached.

A second observation is that the MAP adaptation factor improves the performance as long as it is smaller than 1. This result shows that adaptation from the UBM is important. However, a full adaptation ($\alpha = 0$) leads to a decrease in the performance, which indicates that the general information of all classes (available in the UBM) is also very important for the recognition.

In Table 2, optimal parameter values obtained on the validation set are given. The first column shows the type of classifier, the second column describes the meta parameters and the third column shows the classification rate on the validation set. The best combination method gives a significantly higher performance (70.98%) than the best individual classifer, i.e., the on-line system with 69.98%.

Table 3 shows the classification results for the gender classification task on the test set. The best combination achieved a performance of 67.57%. This result is significantly higher than the classification of the best individual classifier (using a z-test with a significance level of 0.05).

A deeper analysis of the different combination strategies is given in Table 4. As can be seen, the combination based on the Average Rule achieves the best performance and is significantly better than the other two methods (z-test, significance level 0.05). Using the Minimum Rule only leads to a small increase of the performance (not significant) and the Maximum Rule surprisingly leads to a performance decrease, compared to the on-line recognizer.

[2] http://www.iam.unibe.ch/~fki/iamondb/.

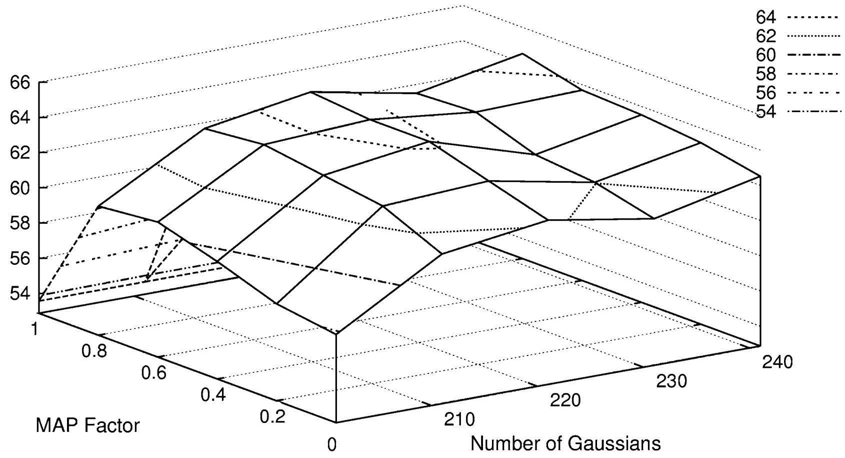


Fig. 4 Optimizing the number of Gaussian mixtures and the MAP adaptation factor of the on-line system on the validation set. Note that the y-axis denotes the classification rate. The *dashed lines* indicate levels of performance, with the maximum occurring at (0.8,220)

Table 1 Gender classification rates on the validation set near the local maximum

α	G		
	201	220	230
0.4	63.35	64.18	62.35
0.6	63.85	64.18	63.52
0.8	63.52	64.51	63.35

Table 2 Gender classification rates on the validation set

Classifier	Meta parameters	Classification rate (%)
On-line	G: 231, α:0.8, φ:0.009	69.98
Off-line	G: 221, α:0.4, φ:0.001	57.88
Combination	G: 231, α:0.6, φ:0.011	70.98

Table 3 Gender classification rates on the test set

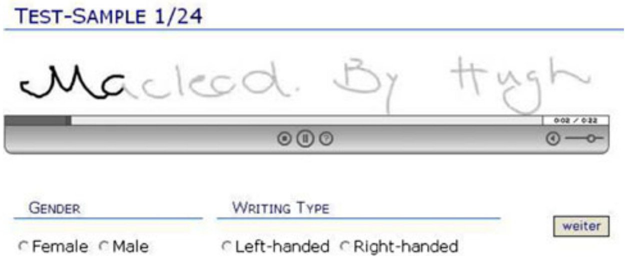
Classifier	Classification rate (%)
On-line	64.25
Off-line	55.39
Combination	67.57

Table 4 Gender classification rates of several combination approaches

Classifier combination	Classification rate (%)
Average Rule	67.57
Maximum Rule	61.93
Minimum Rule	65.58

This may be due to the fact that the likelihoods are normalized before combination.

To compare the performance of the classifiers to that of humans, we asked 20 persons to classify 24 movies of handwriting from different writers each. The persons who did the classification were volunteers from the University of Bern and the German Research Center for Artificial Intelligence (DFKI) at Kaiserslautern, Germany. They were no experts in forensics. The group consisted of 5 female and 15 male participants aged between 23 and 45.


TEST-SAMPLE 1/24

Macleod. By Hugh

Fig. 5 Screen shot of the web interface for human classification

The movies show on-line handwriting from the test set. For "training" purposes, the test subjects also had classified images from other writers available. A screen shot of the web interface is given in Fig. 5. The movie can be viewed with standard Flash or Quicktime plugins. Navigation is possible to each position of the writing. Below the movie, the human can indicate his decision and submit the answers.[3]

The average classification rate of the humans is about 63.88% for the gender recognition task. This performance is lower than the performance of the automatic system. However, no direct comparison can be made because the test set for the humans contains 24 lines, while the test set for the automatic system contains data from 50 writers.

6 Conclusions and future work

In this paper, we have presented a system that classifies the writers' gender from handwritten text. The data is originally given in on-line format and we extract two feature

[3] The test is available under http://www.iam.unibe.ch/~smueller/.

sets, a set of 29 on-line features, and a set of 9 off-line features obtained after converting the on-line data to the off-line format. For the classification, we use Gaussian Mixture Models (GMMs).

In our experiments, the classifier working with the on-line features outperforms the off-line classifier. Combining both, the on-line and the off-line classifier, leads to a significant improvement to 67.57%. These classification results are higher than human classification results. The GMM results for gender classification are similar to results reported on more constrained data in [4].

Acknowledgments This work was supported by the Swiss National Science Foundation program "Interactive Multimodal Information Management (IM)2" in the Individual Project "Visual/Video Processing", as part of NCCR. Special thanks go to Petra and Michael Liwicki, who provided assistance in getting references from the Humboldt-University, Berlin. Furthermore, we thank all volunteers who participated in the on-line classification test.

References

1. Bandi K, Srihari SN (2005) Writer demographic classification using bagging and boosting. In: Proceedings of the 12th international graphonomics society conference, pp 133–137
2. Beech JR, Mackintosh IC (2005) Do differences in sex hormones affect handwriting style? Evidence from digit ratio and sex role identity as determinants of the sex of handwriting. Pers Individ Dif 39(2):459–468
3. Broom ME, Thompson B, et al. (1929) Sex differences in handwriting. J Appl Psychol 13:159–166
4. Cha S-H, Srihari SN (2001) Apriori algorithm for sub-category classification analysis of handwriting. In: Proceedings of the 6th international conference on document analysis and recognition, pp 1022–1025
5. Czyz J, Kittler J, Vandendorpe L (2004) Multiple classifier combination for face-based identity verification. Pattern Recognit 37(7):1459–1469
6. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39(1):1–38
7. Hamid S, Loewenthal KM (1996) Inferring gender from handwriting in Urdu and English. J Social Psychol 136(6):778–782
8. Hattori T, Izumi T, Kitajima H, Yamasaki T (2004) KANSEI information extraction from character patterns using a modified Fourier transform. In: Proceedings of the Sino-Japan symposium on KANSEI & artificial life, pp 36–39
9. Hecker MR (1996) Die Untersuchung der Geschlechtsspezifität der Handschrift mittels Rechnergestützter Merkmalsextraktionsverfahren. PhD thesis, Humboldt-University, Berlin
10. Huber RA (1999) Handwriting identification: facts and fundamentals. CRC Press, Boca Raton
11. Jaeger S, Manke S, Reichert J, Waibel A (2001) Online handwriting recognition: the NPen++ recognizer. Int J Doc Anal Recognit 3(3):169–180
12. Jiang X, Chen Y-F (2008) Facial image processing. In: Bunke H, Kandel A, Last M (eds.), Applied pattern recognition. Springer, Berlin
13. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley, London
14. Liwicki M, Bunke H (2005) IAM-OnDB - an on-line English sentence database acquired from handwritten text on a whiteboard. In: Proceedings of the 8th internetional conference on document analysis and recognition, vol 2, pp 956–961
15. Liwicki M, Schlapbach A, Bunke H, Bengio S, Mariéthoz J, Richiardi J (2006) Writer identification for smart meeting room systems. In: Proceedings of the 7th IAPR workshop on document analysis systems, vol 3872 of LNCS. Springer, pp 186–195
16. Melin H, Koolwaaij JW, Lindberg J, Bimbot F (1998) A comparative evaluation of variance flooring techniques in HMM-based speaker verification. In: Proceedings of the 5th international conference on spoken language processing, pp 2379–2382
17. Newhall SM (1926) Sex differences in handwriting. J Appl Psychol 10:151–161
18. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted gaussian mixture models. Digit Signal Process 10:19–41
19. Richiardi J, Ketabdar H, Drygajlo A (2005) Local and global feature selection for on-line signature verification. In: Proceedings of the 8th international conference on document analysis and recognition, pp 625–629
20. Scheidat T, Wolf F, Vielhauer C (2006) Analyzing handwriting biometrics in metadata context. In: Proceedings of the 8th SPIE conference on the security, steganography, and watermarking of multimedia contents, vol 6072, pp 182–193
21. Schlapbach A, Bunke H (2008) Off-line writer identification and verification using Gaussian mixture models. In: Marinai S, Fujisawa H (eds.), Machine learning in document analysis and recognition, vol 11. Springer, Berlin, pp. 409–428
22. Schlapbach A, Liwicki M, Bunke H (2008) A writer identification system for on-line whiteboard data. Pattern Recognit 41:2381–2397
23. Tenwolde H (1934) More on sex differences in handwriting. J Appl Psychol 18:705–710
24. Wiskott L, Fellous J-M, Krüger N, von der Malsburg C (1995) Face recognition and gender determination. In: Proceedings of the international workshop on automatic face- and gesture-recognition, pp 92–97
25. Wu B, Ai H, Huang C (2003) Audio- and video-based biometric person authentication, vol 2688 of LNCS, chapter LUT-based adaboost for gender classification. Springer, Berlin, pp 104–110