A New Text-Independent GMM Writer Identification System Applied to Arabic Handwriting

Fouad Slimane, Volker Märgner
Institute for Communications Technology (IfN)
Technische Universität Braunschweig
Braunschweig, Germany
Email: {slimane,maergner}@ifn.ing.tu-bs.de

***Abstract*—This paper proposes a system for text-independent writer identification based on Arabic handwriting using only 21 features. Gaussian Mixture Models (GMMs) are used as the core of the system. GMMs provide a powerful representation of the distribution of features extracted using a fixed-length sliding window from the text lines and words of a writer. For each writer a GMM is built and trained using words and text lines images of that writer. At the recognition phase, the system returns log-likelihood scores. The GMM model(s) with the highest score(s) is (are) selected depending if the score is computed in Top-1 or Top-n level. Experiments using only word and text line images from the freely available Arabic Handwritten Text Images Database written by Multiple Writers (AHTID/MW) demonstrate a good performance for the Top-1, Top-2, Top-5 and Top-10 results.**

***Keywords*-Writer Identification; Arabic Text; AHTID/MW database; Gaussian Mixture Models;**

I. Introduction

This work focuses on the identification of writers using Arabic word and text line images. Two concepts are considered to be critical to writer identification: no one person writes exactly the same way twice, and no two people write exactly alike. These two principles, although oversimplified and disputable, clearly highlight two factors that directly conflict when attempting to identify a person based on one or few handwritten words [1]. Figure 1 shows examples of an Arabic word written by different writers to show the differences from writer to writer. Our goal in this work was to automate the process of writer identification using word and text line scanned images of handwriting.

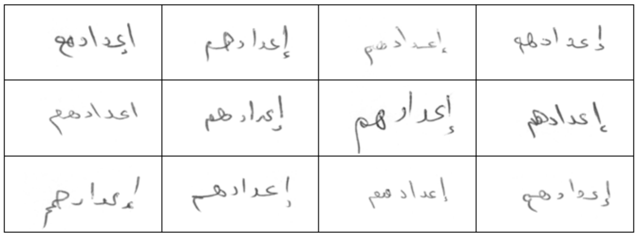

Figure 1. Example of an Arabic word written by twelve different writers

Writer identification is used in biometric and forensic applications, in which the writer can be identified based on a word, a text line or a document. The writer identification of handwritten documents has great importance in the criminal justice system and has been widely explored in forensic handwriting analysis. The relationships between characters, the shape and style of writing will vary from one person to another. However, handwriting is a personal skill with individual characteristics. Therefore, it can be very challenging to determine the best method for correctly identifying a writer.

Writer identification can be divided into two categories: text independent and text dependent. Text-independent writer identification systems can work on any given text, whereas text-dependent writer identification systems require certain known text to be written. In this work, text-independent writer identification of offline Arabic handwritten text is addressed. We present the writer identification problem using scanned Arabic handwritten word and text line images. The objective is to identify the writer of one word or one line of handwritten text (two or more words). In this paper, we propose a robust Gaussian Mixture Models (GMMs) based method, in the framework of the a priori approach, for writer identification of Arabic word images. The method is based on word and text line image modeling with a sequence of only 21 features extracted from a sliding window. A GMM can be viewed as a single-state HMM with a Gaussian mixture observation density [2]. Using GMMs instead of HMMs for writer identification present many advantages such as: (1) text lines, words or characters can be modeled using GMMs, because each writer is represented by only one model; (2) GMM are less complex than HMM because we use only one state for each model; (3) We don't need any transcription during the training; etc.

The rest of this paper is organized as follows. Section II summarizes previous publications related to writer identification. Section III describes the proposed system for writer identification. Section IV is dedicated to the experimental results and it is followed by a conclusion and future work.

2167-6445/14 $31.00 © 2014 IEEE
DOI 10.1109/ICFHR.2014.124

II. Related Work

Various methods and approaches have been proposed for writer identification [3], [4], [5].

A HMM based approach for writer identification and verification was proposed by Schlapbach et al. [6]. For each writer, they build an individual recognizer and train it on text lines of that writer. Said et al. proposed a system for writer identification using textural features derived from the grey-level co-occurrence matrix and Gabor filters [7]. Whole pages of handwritten text are used for this method. Cha et al. propose a writer verification system [8]. This system takes two pages of handwritten text as input and determines if they were written by the same writer. The used features extracted from a text page include character height, stroke width, frequency of loops and blobs, and writing slant and skew.

Somaya Al-Ma'adeed et al. presented a text-dependent writer identification method in Arabic using only 16 words [1]. The extracted features include some edge-based directional features such as height, area, length, and three edge-direction distributions with different sizes. Wighted Euclidean distance has been used as classifier. The used database contains 32,000 Arabic text images from 100 people. They did not mention the Top-1 accuracy of the method, but the best result in Top-10 was 90% when 3 words were used. This method used edge-based directional probability distributions, combined with moment invariants and structural word features, such as length, height, area, length from baseline to lower edge and length from baseline to upper edge.

Awaida and Mahmoud propose a writer identification of handwritten Arabic text [5]. Several types of structural and statistical features were extracted from Arabic handwritten text such as connected component features, gradient distribution features, windowed gradient distribution features, contour chain code distribution features, and windowed contour chain code distribution features. A nearest neighbor classifier was used with the Euclidean distance measure. Data reduction algorithms (viz. principal component analysis, linear discriminant analysis, multiple discriminant analysis, multidimensional scaling, and forward/backward feature selection algorithm) were used. A database of 500 paragraphs handwritten in Arabic by 250 writers was used. The paragraphs used were randomly generated from a large corpus. The nearest neighbor system provided the best accuracy in text-independent writer identification with the 250 writers and with the backward feature selection algorithm (using 54 out of 83 features). The system attained a Top-1 result of 75.0%, top-5 result of 91.8%, and Top-10 result of 95.4%.

On one hand, it is worth noting that only few published works present writer identification systems based on words and text lines levels. Most of them work with paragraphs or pages. To overcome some of these limitations, our research work applies writer identification on Arabic handwritten text of word and text lines from 53 different writers. On the other hand, most of published systems use a lot of features. The proposed system is built using only 21 features with a simple and fast GMM based classifier.

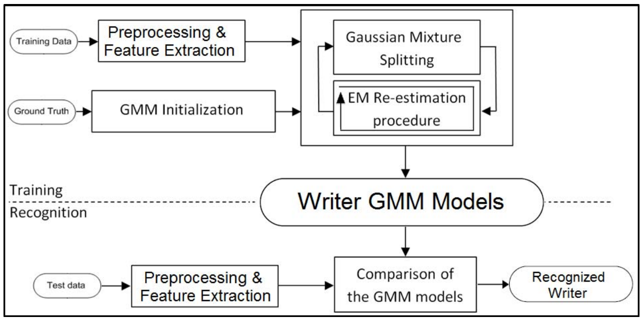
Training Data
Preprocessing & Feature Extraction
Gaussian Mixture Splitting
EM Re-estimation procedure
Ground Truth
GMM Initialization
Training
Recognition
Writer GMM Models
Test data
Preprocessing & Feature Extraction
Comparison of the GMM models
Recognized Writer

Figure 2. GMM based writer identification system

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. The second one computes likelihood estimators of each writer model. For each writer, a GMM is built and trained with data coming from that writer only. To train the GMM, a set of features are extracted from a word or text line image using the sliding window technique. GMM models are trained using the Expectation Maximization (EM) algorithm [9]. The recognition is performed through a simple score comparison of the trained GMM models. The CPU cost of our approach increases linearly with the number of writers, and increases also linearly with the width of the words and text lines.

A. Preprocessing and Feature Extraction

All images of word and text lines in the used database are scanned in gray level. Different pens have been used to write the text lines. In order to eliminate the effect of the pen on the writer identification rate, all used images are binarized using an adapted algorithm of the well known Fischer method [10]. We remove also the white row pixels in top and in bottom of word and text line images if they exist.

In the next step, we extract features using the sliding window technique. The window moves from right to left. The window width of the sliding window and the shift pixels were optimized in an independent validation experiment consisting of word images from the 53 writers. The highest writer identification rate was achieved using a window width of 25 pixels and a shift of one pixel. This window width was used in all experiments to extract the features from word and text line images.

Some features are computed using the lower and upper baselines. These baselines divide the image into three zones as illustrated in Figure 3:

- **Core zone** of a word, that is the zone bounded by the two upper and lower baselines, does not contain ascenders and descenders
- **Upper zone** of a word, that is the zone above the upper baseline where ascenders can be found
- **Lower zone** of a word, that is the zone under the lower baseline where descenders can be found

The lower baseline is detected based on the horizontal projection according to the longest peak, and the upper baseline is detected according to the longest peak above the lower baseline. Our algorithm is based on the one described in [11].

For every window of pixels, twenty-one 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:

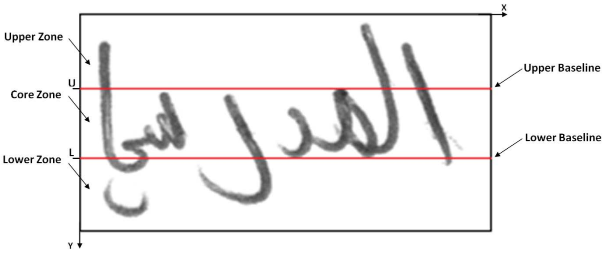

Figure 3. Division of word image in three zones using the upper and lower baselines

- Density of pixels in the window
- Vertical position of the gravity center in the whole window (W) with respect to the lower baseline. The result is normalized by the height h of the window as presented in Equation 1.

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

where L is the position of the lower baseline.
- Mean of 12 Zernike moments computed from the window [12]. 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 * denotes the complex conjugate.

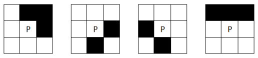

Figure 4. The used typological masks around a background pixel P

- Mean of vertical projection normalized by the window width
- Mean of horizontal projection
- Standard deviation of vertical projection normalized by the window width
- Standard deviation of horizontal projection
- Derivate of vertical projection vector profile
- Derivate of horizontal projection vector profile
- Mean of vertical runs
- Mean of horizontal runs
- Standard deviation of vertical runs
- Standard deviation of horizontal runs
- Number of white pixels according to each of the four typological masks shown in Figure 4 in the window normalized by the size of the window (4 features)
- Number of white pixels according to each of the four typological Masks shown in Figure 4 over upper baseline normalized by the size of the upper zone (4 features). The upper zone of a word is the zone above the upper baseline where ascenders can be found.

Using the sliding window technique, no segmentation into letters is made and the word or text line image is transformed into a sequence of feature vectors where the number of rows corresponds to the number of components of each feature vector, and the number of columns is equal to the number of analysis windows.

The sequence of twenty-one-dimensional feature vectors thus obtained from each word or line of text image is used to train the GMMs. As a result of the training procedure, we obtain for each writer a GMM that is specially adapted to the individual handwriting style of that writer.

B. Gaussian Mixture Models for Writer Identification Modeling

Gaussian Mixture Models (GMMs) are used to model the handwriting of each person of the underlying population. We modeled the distribution of the feature vectors using Gaussian mixture density.

For a D-dimensional feature vector x the mixture density for a specific 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 $D \times D$ covarience matrix $\sum_i$. The parameters of a writer's density model are denoted 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 coefficients of the feature vectors 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].

During training, the iterative EM algorithm is used to refine the GMM parameters (component weights, means and variances) to monotonically increase the likelihood of the estimated model for the observed feature vectors [14]. In our experiments, we used the EM algorithm to build the models by applying a simple 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 4096 Gaussians as reference for our writer 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:

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

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

During decoding, a word or text line to be classified is presented to the GMM of each writer. Each GMM outputs the log-likelihood score and the standard deviation for the given text line or word image. The log-likelihood scores are sorted in decreasing order and the text line is assigned to the best ranked writer. In this work, we don't implement the rejection mechanism. In applications where a wrong identification implies a high cost, the rejection mechanism can be implemented easily as follows: compute the difference between the log-likelihood of the first and the second best ranked writer, normalized by the length of the word or text line. Then, if this measure is above a certain threshold, assign the word or text line to the best ranked writer. Otherwise, no decision about the identity of the word or text line is made. This method is inspired from [6].

System performances are evaluated in terms of writer identification rates using an unseen set of word and text line images.

Our GMM-based system is implemented using the Hidden Markov Model Toolkit (known as HTK Toolkit)[1] [15].

IV. Experimental Results

To evaluate the performance of our writer identification system, experiments have been conducted on the freely available AHTID/MW Database [16]. In all tests, identification rates have been evaluated at word and text line level independently to their length.

A. The AHTID/MW Database

The Arabic Handwritten Text Images Database written by Multiple Writers (AHTID/MW) has been built at the MIRACL Lab, ISIMS, University of Sfax - Tunisia in joint collaboration with the Institute for Communications Technology (IfN), Braunschweig - Germany. It can be used for research in the recognition of Arabic handwritten text, word segmentation, word spotting and writer identification. The AHTID/MW contains 3710 text lines and 22,896 words written by 53 native writers of Arabic with different ages and educational backgrounds with no restrictions for choosing the type of pen. A variety of Arabic handwriting styles are presented in the AHTID/MW dataset as illustrated by Figure 5. These images are divided into five equally large sets. The four first sets are available for scientific community and the fifth set is kept internal for potential future evaluation of systems in blind mode. Each word/text line image in AHTID/MW database is fully described using an XML file containing ground truth information. The database is freely available for researchers worldwide. Actually, AHTID/MW database is used by many groups all over the world working on Arabic handwritten text recognition. For more details, we refer to [16] and the *IAPR-TC11 Datasets List* web page[2].

To compare writer identification systems developed by different groups, a first competition was held at ICFHR'2014. The scientific objectives of that competition are to measure the capacity of recognition systems to identify the writer using word, text line and paragraph images. The main difficulty is probably in the similarity between the writing styles, the quality of the images as they are scanned on grey level and in the possibility to recognize the writer using one word, one line or one paragraph image[3].

B. Writer Identification Results and Discussion

The experiments are based on word and text line images. For each writer we use about 71 text lines (2/3 text line for training and 1/3 for testing). We perform full four-fold cross validation experiments using the four sets of the AHTID/MW database. The first three sets are used for training and the fourth set is used for testing.

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

[2] http://www.iapr-tc11.org/mediawiki/index.php/Datasets_List

[3] http://www.ifn.ing.tu-bs.de/news/icfhr2014/

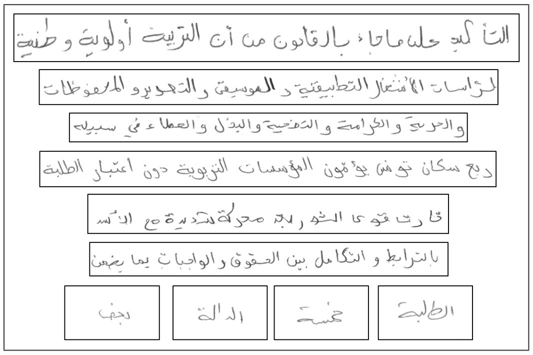

Figure 5. Arabic handwritten word and text line images from AHTID/MW database written by different writers

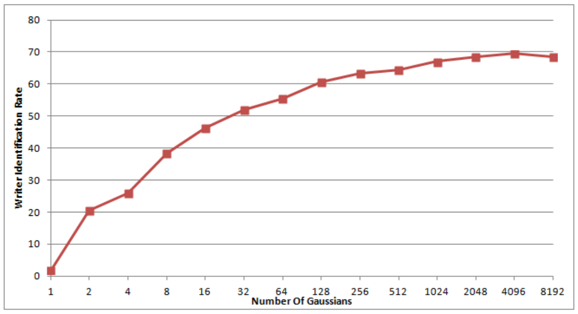

Figure 6. Top-1 writer identification rate as a function of the number of Gaussians in the writer model with GMMs.

In a first set of writer identification experiments, we tried to find the optimal number of Gaussians for the writer models. We tested our system with 1, 2, 4, 8, ... , 4096, 8192 Gaussians in the mixture, doubling the number of Gaussians after each 10 iterations of the EM algorithm. Figure 6 illustrates the evolution of the writer identification rates as a function of the number of Gaussians in the models. The Top-1 highest writer identification rate of 69.48 is achieved using 4096 Gaussians.

In Table I, an n-best list is shown, where the writer identification rate based on the first n ranks is presented. The system provided the best accuracy in text-independent writer identification for all 53 writers with a Top-1 result of 69.48%, Top-2 result of 77.6%, Top-5 result of 85.24%, and Top-10 result of 92.01% using text line images and 23.03% in Top-1, 34.22% in Top-2, 49.84% in Top-5 and 63.99% in Top-10 for word images.

Table I
TOP-N WRITER IDENTIFICATION RESULTS USING GMMS-BASED SYSTEM

Writer identification rate	Top-1	Top-2	Top-5	Top-10
Line level (%)	69.48	77.36	85.24	92.01
Word level (%)	23.03	34.22	49.84	63.99

We compare the GMM based system to a continuous Hidden Markov Models (HMMs) based system using the same features and the same parameters of the sliding window technique. After many tests, each writer HMM consists of an optimal number of 3 states. The output probability distribution is modeled by a mixture of Gaussian components. In this experiment, the optimal number of mixture components is 1024 to reach the best writer identification rate. The best writer identification is about 69.7% in Top-1, 77.91% in Top-2, 85.46% in Top-5 and 92.01% in Top-10 for text line images and 23.04% in Top-1, 32.09% in Top-2, 48.06% in Top-5 and 62.74% in Top-10 for word images.

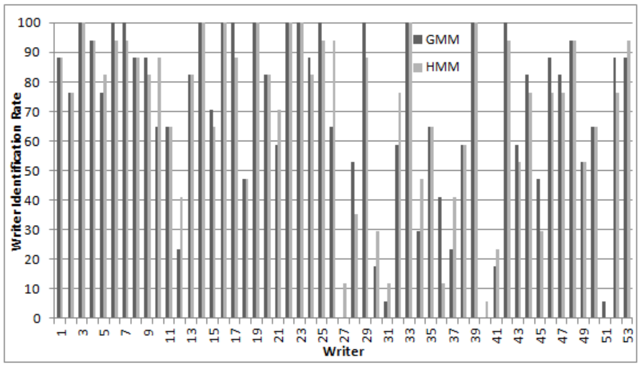

Figure 7. Top-1 text line identification rates for each writer using GMM and HMM based systems

Figures 7 and 8 show the results in detail for each writer using GMM and HMM based systems respectively for Top-1 text line and Top-10 word images. As we can see in Figure 7, we reach 100% accuracy with 14 writers and less than 50% with 12 writers in Top-1 level using the GMM based system and respectively 8 with 100% and 13 with less than 50% using the HMM based system. We recognize with more than 80% 11 writers in Top-10 using only one word image using the GMM based system and respectively 16 writers using the HMM based system as illustrated in Figure 8. As a consequence, we interpret from results that writer identification using only one word which can be a proposition (one character) is more complex than using one text line.

The results are comparable but the HMMs are conceptually more complex than GMMs (GMMs can be seen as one-state Hidden Markov Models and one output distribution function, which leads to significantly shorter training times). In GMMs also only the parameters of the output distribution function have to be estimated during training compared to HMMs where the state transition probabilities have to be estimated as well.

To the best of our knowledge, this is the first published work on writer identification using the AHTID/MW database. A detailed comparison between the proposed GMM based system and the participant systems to the ICFHR'2014 writer identification competition are presented

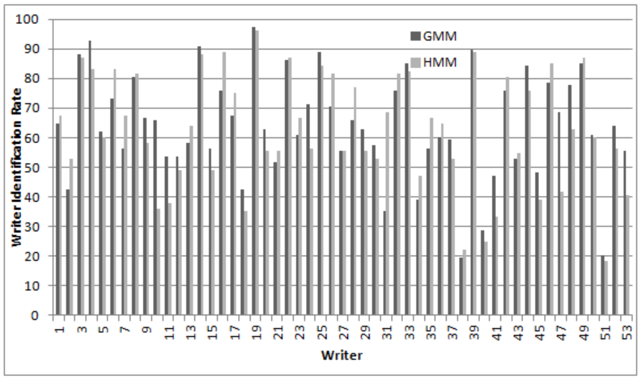

Figure 8. Top-10 word identification rates for each writer using GMM and HMM based systems

in the competition paper.

V. Conclusion and future work

In this paper we use Gaussian mixture models to address the task of text-independent writer identification using one word or one text line image. We model the handwriting distribution for writers using GMMs and only 21 local features. When presented with a word or text line 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 writer. A database of 3710 Arabic handwritten text lines written by 53 writers was used in the analysis and experiments. The accuracy results for different numbers of writers were presented and analyzed. Our GMMs-based system provided the best accuracy in text-independent writer identification for all 53 writers with a Top-10 result of 63.99% for word images and 92.01% for text line images. We reach 100% accuracy using text lines for 14 writers in Top-1.

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

Acknowledgment

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

References

[1] S. Al-Maadeed, "Text-dependent writer identification for arabic handwriting," *JECE*, vol. 2012, pp. 13:13–13:13, 2012.

[2] A. Schlapbach and H. Bunke, "Off-linewriter identification using gaussian mixture models," *International Conference on Pattern Recognition (ICPR)*, vol. 3, pp. 992–995, 2006.

[3] T. A. Abu-Ain, W. A. Abu-Ain, S. N. H. S. Abdullah, and K. Omar, "Off-line arabic character-based writer identification a survey," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 1, no. 2, pp. 161–166, 2011.

[4] Sreeraj.M and S. M. Idicula, "Article: A survey on writer identification schemes," *International Journal of Computer Applications*, vol. 26, no. 2, pp. 23–33, July 2011.

[5] S. M. Awaida and S. A. Mahmoud, "Writer identification of arabic text using statistical and structural features," *Cybernetics and Systems*, vol. 44, no. 1, pp. 57–76, 2013.

[6] A. Schlapbach and H. Bunke, "Using hmm based recognizers for writer identification and verification," *International Workshop on Frontiers in Handwriting Recognition (ICFHR)*, pp. 167–172, 2004.

[7] H. Said, T. Tan, and K. Baker, "Personal identification based on handwriting," *Pattern Recognition*, vol. 33, no. 1, pp. 149 – 160, 2000.

[8] S. hyuk Cha, S. N. Srihari, Sargur, and N. Srihari, "Multiple feature integration for writer verification," *Proceedings of the Seventh Workshop on Frontiers in Handwriting Recognition*, pp. 333–342, 2000.

[9] 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.

[10] R. Fisher, S. Perkins, A. Walker, and E. Wolfart, "Adaptive thresholding," *http://homepages.inf.ed.ac.uk/rbf/HIPR2/adpthrsh.htm*, 2003.

[11] R. El-Hajj, L. Likforman-Sulem, and C. Mokbel, "Arabic handwriting recognition using baseline dependant features and hidden markov modeling," *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 893–897, 2005.

[12] 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.

[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] 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.

[15] 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.

[16] A. Mezghani, S. Kanoun, M. Khemakhem, and H. Abed, "A database for arabic handwritten text image recognition and writer identification," *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 399–402, 2012.