Evaluation of Texture Features for Offline Arabic Writer Identification

Chawki Djeddi[1], Labiba-Souici Meslati[2], Imran Siddiqi[3], Abdelllatif Ennaji[4], Haikal El Abed[5], Abdeljalil Gattal[1]

[1] LAMIS Laboratory, University of Tebessa, Tebessa, Algeria
[2] LISCO Laboratory, Badji Mokhtar-Annaba University, Annaba, Algeria
[3] Department of Computer Science, Bahria University, Islamabad, Pakistan
[4] LITIS Laboratory, Rouen University, Rouen, France
[5] Technical Trainers College, German International Cooperation, Riyadh, Kingdom of Saudi Arabia

c.djeddi@mail.univ-tebessa.dz

***Abstract*— Biometric identification of persons has mainly been based on fingerprints, face, iris and other similar attributes. We propose a handwriting-based biometric identification system using a large database of Arabic handwritten documents. The system first extracts, from each handwritten sample, a set of features including run lengths, edge-hinge and edge-direction features. These features are used by a Multiclass SVM (Support Vector Machine) classifier. Experiments are conducted on a new large database of Arabic handwritings contributed by 1000 writers. The highest identification rate achieved by the combination of run-length and edge-hinge features stands at 84.10%.**

Keywords*— *Arabic handwriting; KHATT database; Offline handwriting; Writer identification; Textural features.

I. INTRODUCTION

The last few years have witnessed a significant increase of research in different areas of biometrics. Notable advancements in this area have resulted in many biometric modalities such as retina recognition, vein pattern recognition, ear geometry, facial thermography, palm print recognition, DNA recognition, speaker recognition, handwritten signatures, keystroke dynamics and gait recognition. Biometric identification modalities can be divided into physical and behavioral biometrics. Physical biometrics employs some physical characteristics to identify an individual. Behavioral biometrics is based on the behavioral traits learnt and acquired over time.

Over the recent years, research in analysis and recognition of handwriting has primarily focused on the use of handwriting as biometric identifier and, a number of signature recognition systems [01, 02, 03] have been proposed. The work of Srihari et al. [04] is seen as the first serious attempt to prove and exploit the individuality of handwriting to develop a system using handwritten documents to identify individuals. The most recent studies by Bulacu et al. [06] and Siddiqi et al. [05] have shown that individuals can be effectively identified using their handwritten samples.

Like other biometric modalities, writer recognition systems can be used in two different modes, identification mode and verification mode. In identification mode, the goal is to determine which writer, amongst a set of known writers, has written a given sample. In verification, the objective is finding out whether two handwritten documents are written by the same person or not. In other words, writer identification is a one-to-many search while writer verification is a one-to-one problem. Writer recognition approaches can be categorized into two distinct families: text-dependent approaches and text-independent approaches. In text-dependent methods, the writer must write exactly the same predefined or given text. Text-independent writer recognition is the process of identifying or verifying the identity of the writer without any constraints on the textual content of the samples being compared.

In this work, a new text-independent biometric personal identification method using Arabic handwritten documents is introduced. The proposed method consists of two main stages: feature extraction and classification (writer identification). In the first step, run-lengths on white and black pixels, edge-hinge and edge direction features are extracted from handwritten documents. In classification, the features of the query document are matched with the features corresponding to handwritten documents in the database. In our study, we have used Multi-class SVM (Support Vector Machine) with one against all method as classifier. The experiments are conducted on a new publicly available database KHATT [07] comprising documents written by 1000 different writers.

The rest of this paper is organized as follows. Section 2 outlines the notable contributions in Arabic writer identification. The database used is presented in section 3. In section 4, we present the feature extraction method followed by a summary of obtained results. We then present a detailed analysis of these results and finally conclude our findings in the last section.

II. ARABIC WRITER IDENTIFICATION

Writer identification from Arabic handwritten documents has not been addressed as extensively as writer identification from documents in other scripts like Latin or Chinese. The past few years, however, have seen significant contributions in this area. Al-Zoubeidy et al. [08] proposed the use of multi-channel Gabor filtering and gray scale co-occurrence matrices to characterize the writing style of writers. Gazzah et al. [09] combined a set of global and local features. The

978-1-4799-3243-6/14 $31.00 © 2014 IEEE
DOI 10.1109/DAS.2014.76

global features are based on 2D Discrete Wavelet transform whereas the local features capture information like slant, line height, ascenders and descenders etc.

Al-Dmour et al. [10] developed a texture based method for writer characterization and extracted a set of features based on spectral-statistical measures (SSMs) of texture. Bulacu et al. [11] proposed a combination of textural and allographic features and achieved significantly improved performance as compared to that of individual features.

Abdi et al. [12] represented each word by the minimum perimeter polygon (MPP) contours and extracted a set of six features mainly including distributions of lengths, directions and curvatures etc.

Authors in [13], characterize the author of a given sample by computing the fractal dimension of writing using the box counting method. The box counting based fractal dimension is complemented by multi-fractal dimensions which are calculated using the Diffusion Limited Aggregates (DLA).

Al-Ma'adeed et al. [15] employed moment invariants and edge-based directional probability distributions as features to characterize the writer. These features are complemented by local features including area, length, height, lengths from baseline to upper and lower edges etc. computed from individual words.

Chen et al. [16] evaluated the effectiveness of removing ruling lines from handwritten documents prior to performing writer recognition. The series of experiments carried out by the authors reveal that removing these lines significantly improves the identification results.

III. DATABASE

Our study is based on handwriting samples extracted from a new publicly available database KHATT [07]. This database contains handwritten Arabic text images and its ground-truth developed to promote research in areas like writer identification, line segmentation, binarization and noise removal besides handwritten text recognition.

The KHATT database contains 4000 grayscale paragraph images containing Arabic texts scanned at different resolutions (200, 300 and 600 dpi) written by 1000 writers of different ages and backgrounds from 18 different countries. Out of the 1000 writers, 677 individuals were male while 323 were female and, 928 were right handed while 72 were left handed. 2000 of these images contain similar text each covering all Arabic characters and shapes whereas the remaining 2000 images contain free texts written by the writers on any topic of their choice. For our experiments, we use the 2000 paragraph images with free text to train the system. From the 2000 paragraph images that contain similar texts, we have chosen 1000 images (of 1000 different writers) to test the proposed system. Figure 1 shows two paragraph images with different textual content written by the same person while Figure 2 presents two other paragraph images written by the same person but containing the same text.

IV. FEATURE EXTRACTION

To characterize the writing style of an individual, we have implemented three methods, these include: run-length distribution [17], edge-hinge distribution [06] and edge-direction features [06]. Below is a brief description of each of these three features.

A. Run-Length features

The run-length features are computed on a binary image taking into consideration the black and white pixels that correspond to the ink trace and to the background of the text images respectively [17]. The proposed method considers horizontal (f1), vertical (f2), left-diagonal (f3) and right-diagonal (f4) run-lengths on black extracted from the image of handwriting as well as the horizontal (f5), vertical (f6), left-diagonal (f7) and right-diagonal (f8) run-lengths on white.

We have tested these run-length features in the ICDAR 2011 Writer Identification Contest [18], ICDAR 2011 Arabic Writer Identification Contest [19], ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification [20], ICFHR 2012 Competition on Writer Identification-Challenge 1: Latin/Greek Documents [21] and ICFHR 2012 Competition on Writer Identification - Challenge 1: Arabic Scripts [22]. These features realized interesting results in these competitions.

Figure 1. Two paragraph images of different content written by the same person.

Figure 2. Two paragraph images containing similar text written by the same person.

B. Edge-Direction features

Edge direction features, proposed by Bulacu et al. in 2003 [6] have been effectively used for statistical analysis of online as well as offline handwritings. To calculate the edge direction features, two steps are necessary. The first step is to refine the writing by performing a convolution by two orthogonal kernels (Sobel) to increase the importance of the edges before thresholding the resultant image. Directional analysis then comes through a sliding window of n×n pixels allowing the construction of the distribution function.

Within the window, the possible directions of the edge fragments are analyzed. Whenever the central pixel of a window is on an edge fragment, we determine the direction of the later. The number of the edge fragments found in each direction is then accumulated to constitute the probability density function.

An example of this method is illustrated in Figure 3. In our experiments, we considered fragments of lengths 2 (f13), 3 (f14), 4 (f15) and 5 (f16) pixels. The edge direction feature, in addition to the prevailing direction of the slope of writing, also provides an indication on the regularity of writing. For a more detailed description of the edge direction features, please refer to [06].

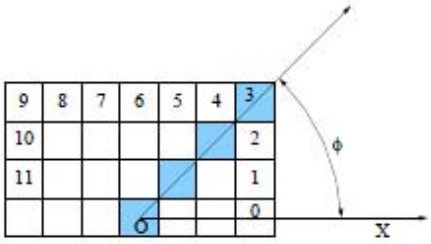

Figure 3. Extraction of edge-direction distribution (Image reproduced from [06]).

C. Edge-hinge features

Edge-hinge distribution is known to be one of the most effective features for characterizing the writing style of writers. The edge-hinge distribution was proposed in 2003 by Bulacu et al. [06]. It characterizes the change of directions in writing and its computation is similar to that of edge distribution presented in section 4.2. For edge-hinge features, instead of analyzing the direction of the edge fragments emerging from the central pixel (if it is a text pixel), we are interested in computing the angle separating the two edge fragments (i.e. connected sequences of pixels). An example of edge hinge extraction is shown in Figure 4. In our experiments, we considered fragments lengths of 4 (f9), 5 (f10), 6 (f11) and 7 pixels (f12). For a more detailed description of the edge hinge features, please refer to the paper [06].

V. Writer Identification

Once the handwritten documents in the training and test set are represented by the set of features, writer identification is carried out using a Multiclass Support Vector Machines (SVM). The chosen documents from the KHATT database [07] are divided into two parts, training set and test set. 67% of the chosen documents are used for training and the rest (33%) are used as a test set. We build a SVM model on the training set and it is used to retrieve the writers' identity of documents in test set. For each query document in the testing set, we not only find the Top-1writer but a longer list up to a given rank (Top-10) thus increasing the chance of finding the correct writer in the retrieved list.

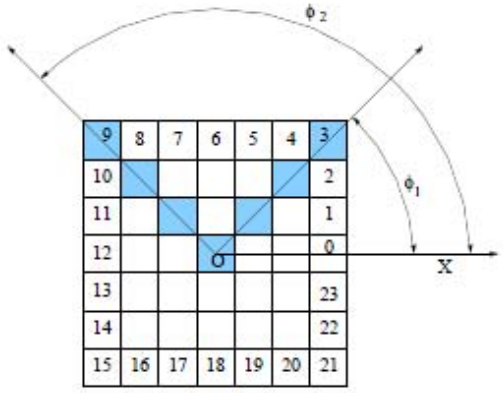

Figure 4. Extraction of edge-hinge distribution (Image reproduced from [06]).

TABLE I. Overview of proposed features and their dimensions.

Feature	Description	Dim
f1	Horizontal run-lengths on black	200
f2	Vertical run-lengths on black	200
f3	Left-diagonal run-lengths on black	200
f4	Right-diagonal run-lengths on black	200
f5	Horizontal run-lengths on white	200
f6	Vertical run-lengths on white	200
f7	Left-diagonal run-lengths on white	200
f8	Right-diagonal run-lengths on white	200
f9	Edge-hinge with fragment length of 4 pixels	1024
f10	Edge-hinge with fragment length of 5 pixels	1600
f11	Edge-hinge with fragment length of 6 pixels	2304
f12	Edge-hinge with fragment length of 7 pixels	3136
f13	Edge-direction using 4 angles	4
f14	Edge-direction using 8 angles	8
f15	Edge-direction using 12 angles	12
f16	Edge-direction using 16 angles	16

VI. Experiments

To evaluate the proposed approach, we conducted two series of experiments: the first one is designed to evaluate the performance of individual features whereas the second aims at testing the effectiveness of different combinations of these features. We report the Top 1, Top 5 and Top 10 identification rates for writer identification task.

For each feature, Table 1 summarizes the corresponding description and the dimension, whereas Table 2 presents the

performance of these features. Although the feature performances vary significantly, it can be noticed that the edge-hinge features (f9-f12) outperform the run-lengths features (f1-f8), with f12 (Edge-hinge with fragment length of 7 pixels) achieving the best results on identification task.

TABLE II. IDENTIFICATION RESULTS ON INDIVIDUAL FEATURES.

Feature	Top 1	Top 5	Top 10
f1	17.00%	32.30%	38.60%
f2	16.40%	33.60%	39.20%
f3	18.30%	35.10%	42.00%
f4	16.00%	33.30%	41.60%
f5	2.40%	6.80%	9.70%
f6	2.90%	6.90%	8.80%
f7	4.80%	10.60%	13.00%
f8	3.90%	8.40%	10.80%
f9	71.00%	85.60%	88.30%
f10	75.60%	88.70%	89.40%
f11	78.20%	90.00%	91.30%
f12	79.30%	91.80%	92.90%
f13	6.20%	18.20%	26.50%
f14	20.00%	39.40%	49.30%
f15	25.80%	47.70%	57.00%
f16	32.80%	52.40%	56.50%

Table 3 summarizes the performance of some of the feature combinations we have tested. For writer identification, the highest rate we achieved stands at 84.10% in Top 1, 91.80% in Top 5 and 92.80% in Top 10 when combining run-lengths on white and black pixels with edge-hinge (fragment length of 4 pixels) distribution (f1-f8, f9).

TABLE III. IDENTIFICATION RESULTS ON FEATURES COMBINATION.

Features combinations	Top 1	Top 5	Top 10
f1, f2, f3, f4	58.20%	71.90%	73.20%
f5, f6, f7, f8	25.80%	42.60%	48.10%
f1, f2, f3, f4, f5, f6, f7, f8	70.60%	82.90%	84.80%
f13,f14,f15,f16	46.00%	61.40%	63.90%
f1, f2, f3, f4, f5, f6, f7, f8, f9	84.10%	91.80%	92.80%
f1, f2, f3, f4, f5, f6, f7, f8, f10	83.90%	91.80%	92.80%
f1, f2, f3, f4, f5, f6, f7, f8, f11	83.50%	92.40%	93.30%
f1, f2, f3, f4, f5, f6, f7, f8, f12	83.90%	92.50%	93.10%
f1, f2, f3, f4, f13,f14,f15,f16	64.90%	78.80%	80.90%

When comparing the identification performances across the three types of features, it can be seen that the identification results are much poor when using run-lengths individually but it is comparable with the edge-hinge features when we combine all the run-lengths features.

Table 4 summarizes the performance of the most recent studies on Arabic writer identification on different databases. Bulacu & al. [11] currently hold the best performance results with 88% in Top 1 and 99% in Top 10 on 350 writers. We have achieved an identification rate of 70.60% in Top 1, 82.90% in Top 5 and 84.80% in Top 10 by using the run-length features and we have improved the results by combining the run-lengths features with edge-hinge features to achieve an identification rate of 84.10% in Top 1, 91.80% in Top 5 and 92.80% in Top 10.

TABLE IV. COMPARISON OF SOME ARABIC WRITER IDENTIFICATION METHODS.

Reference	Writers	Top 1	Top 5	Top 10
[08]	20	92.80%	-	-
[10]	20	90.00%	-	-
[13]	50	-	92.60%	-
[09]	60	95.68%	-	-
[16]	60	74.30%	-	-
[12]	82	90.20%	96.30%	97.50-%
[15]	100	-	-	93.80%
[23]	275	93,53%	98,47%	99,13%
[11]	350	88.00%	-	99.00%
Our method	1000	84.10%	91.80%	92.80%

VII. CONCLUSION

Writer identification is a relatively new biometric modality that has received significant research attention in the recent years. Handwriting biometrics can be used in the forensic applications to identify individuals based on their writing characteristics by comparing unlabeled handwritten texts with labeled handwritten samples. This paper presented a handwriting biometric identification method based on multiclass SVM classifier and a set of run-lengths combined with edge-hinge features extracted from handwritings.

The obtained results demonstrate the effectiveness of the run-length features at modeling handwriting for writer identification. The database used in our experiments contains handwritten documents from 1000 different writers, and the obtained results are promising especially when considering that only 2 documents per writers are used in training phase.

REFERENCES

[1] K. Huang, H. Yan, "Off-line Signature Verification Based on Geometric Feature Extraction and Neural Network Classification", Pattern Recognition, Vol. 30, No. 1, pp. 9-17, 1997.

[2] R. Sabourin, R. Plamondon, "Progress in the Field of Automatic Handwritten Signature Verification Systems Using Gray-level Images", Inter. Workshop on Frontiers in Handwriting Recognition, Abril, Montreal, 1990.

[3] R. Bajaj, S. Chaudhury, "Signature Verification Using Multiple Neural Classifiers", Pattern Recognition, Vol. 30, No. 1, pp. 1-7, 1997

[4] S. Srihari, S. Cha, H. Arora, and S. Lee, "Individuality of Handwriting," J. Forensic Sciences, Vol. 47, No. 4, pp. 1-17, July 2002.

[5] I. Siddiqi and N. Vincent, "Text independent writer recognition using redundant writing patterns with contour- based orientation and curvature features". In Pattern Recognition Journal, 43 (11): 3853-3865, 2010.

[6] M. Bulacu, L. Schomaker, L. Vuurpijl, "Writer identification using edge-based directional features", Proc. of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003), IEEE Computer Society, pp. 937 – 941, vol. II, Edinburgh, Scotland, 2003.

[7] S. A. Mahmoud, I. Ahmad, M. Alshayeb, W. G. Al-Khatib, M. T. Parvez, G. A. Fink, V. Margner, and H. EL Abed, "KHATT: Arabic Offline Handwritten Text Database," In 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, pp. 447-452 , 2012

[8] L. M. Al-Zoubeidy et H. F Al-Najar, “Arabic writer identification for handwriting images”, International Arab Conference on Information Technology, Amman, Jordan, pp. 111-117, 2005.

[9] S. Gazzah, N.E Ben Amara, “Arabic Handwriting Texture Analysis for Writer Identification using the DWT-lifting Scheme“. In 9th ICDAR, vol.2, 1133–1137, 2007.

[10] A. AL-Dmour, R.A Zitar, “Arabic Writer Identification based on Hybrid Spectral-Statistical Measures”. J. Experimental and Theoretical Artificial Intelligence, Vol.19, pp. 307–332, 2007.

[11] M. Bulacu, L. Schomaker, A. Brink, "Text-Independent Writer Identification and Verification on Offline Arabic Handwriting," ICDAR 2007. pp.769-773, 23-26 Sept. 2007

[12] M.N. Abdi, M. Khemakhem, H. Ben-Abdallah, “Off-Line Text-Independent Arabic Writer Identification using Contour-Based Features”. In International Journal of Signal and Image Processing Vol.1, 4–11 (2010).

[13] A. Chaabouni, H. Boubaker, M. Kherallah, A.M. Alimi, H. El Abed, “Fractal and Multi-fractal for Arabic Offline Writer Identification”, International Conference on Pattern Recognition, Istanbul, Turkey, 2010, pp : 1051-4651.

[14] S. M. Awaida, S. A. Mahmoud. “Writer Identification of Arabic Handwritten Digits”, First International Workshop on Frontiers in Arabic Handwriting Recognition, Istanbul, Turkey, 2010.

[15] S. Al-Ma’adeed, E. Mohammed, D. Al Kassis, F. Al-Muslih, “Writer Identification using Edge-Based Directional Probability Distribution Features for Arabic Words”. In: IEEE/ACS International Conference on Computer Systems and Applications, 582–590 (2008).

[16] J. Chen, D. Lopresti, E. Kavallieratou, “The Impact of Ruling Lines on Writer Identification”, in the Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition, Kolkata, India, pp. 439 – 444 , 2010

[17] C. Djeddi, L. Souici-Meslati, “A texture based approach for Arabic Writer Identification and Verification”, IEEE International Conference on Machine and Web Intelligence, ICMWI’2010, Algiers, Algeria, pp: 115-120, 2010.

[18] G. Louloudis, N. Stamatopoulos and B. Gatos, “ICDAR 2011 - Writer Identification Contest”, In Proc of the 11th International Conference on Document Analysis and Recognition, pp. 1475-1479, China, 2011.

[19] A.Hassaine, S. Al-Maadeed, J.M. Alja’am, A. Jaoua and A. Bouridane, “The ICDAR2011 Arabic Writer Identification Contest”, In Proc of the 11th International Conference on Document Analysis and Recognition, pp. 1470-1474, China, 2011.

[20] A. Fornés, A. Dutta, A. Gordo and J. Llados, “The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification”, In Proc. of the 11th International Conference on Document Analysis and Recognition, pp. 1511-1415, China, 2011.

[21] G. Louloudis, B. Gatos and N. Stamatopoulos, “ICFHR2012 Competition on Writer Identification - Challenge 1: Latin/Greek Documents”. In 13th International Conference on Frontiers in Handwriting Recognition (ICFHR'12), pp. 825-830, Bari, Italy, September 2012.

[22] A. Hassaïne and S. Al Maadeed, “ICFHR2012 Competition on Writer Identification - Challenge 2: Arabic Scripts”, 13th International Conference on Frontiers in Handwriting Recognition (ICFHR'12), pp. 831-836, 2012.

[23] C. Djeddi, L. Souici-Meslati and A. Ennaji, “Writer Recognition on Arabic Handwritten Documents”, Image and Signal Processing, Lecture Notes in Computer Science, Volume 7340, pp. 493 – 501, 2012.