Writer Recognition on Arabic Handwritten Documents

Chawki Djeddi[1], Labiba Souici-Meslati[2], and Abdellatif Ennaji[3]

[1] Laboratoire LAMIS, Département de Mathématiques et d'Informatique,
Université de Tébessa, Route de Constantine, 12000, Tébessa, Algérie
[2] Laboratoire LRI, Département d'Informatique,
Université Badji Mokhtar d'Annaba, B.P 12, 23000. Annaba, Algérie
[3] Laboratoire LITIS, UFR des Sciences, Université de Rouen, 76800,
Saint-Etienne du Rouvray, France
c.djeddi@mail.univ-tebessa.dz

Abstract. Recognizing the writer of a handwritten document has been an active research area over the last few years and is at the heart of many applications in biometrics, forensics and historical document analysis. In this paper, we present a novel approach for text-independent writer recognition from Arabic handwritten documents. To characterize the handwriting styles of different writers involved in the evaluation of our approach, we have used two texture methods based on edge hinge features and run-lengths features. The efficiency of the proposed approach is demonstrated experimentally by the classification of 1375 handwritten documents collected from 275 different Arabic writers.

Keywords: writer identification, writer verification, run-lengths, edge hinge, Arabic handwriting.

1 Introduction

Writer recognition based on handwritten documents is a hot and promising research topic in the field of pattern recognition due to its various applications; it is a classical pattern recognition problem [1]. The classification task in pattern recognition is to assign a pattern to one class out of a set of classes. In this paper, a pattern is a sample of handwritten text and a class represents a writer.

Writer recognition is the process of automatically recognizing who is writing on the basis of individual information included in handwritten documents. Writer recognition refers to two different tasks: Writer identification and writer verification. Writer identification determines which writer provides a given handwriting form amongst a set of known writers. Writer verification consists to decide on two handwritten documents and determine if they are written by the same writer or by two different writers.

Writer recognition approaches can be categorized into two distinct families: text-dependent approaches and text-independent approaches: In text-dependent approaches, the writer must write exactly a predefined or a given text. The text-independent writer recognition is a process of identifying or verifying the identity of the writer without constraint on the text content.

Writer recognition systems are involved in many applications such as biometric recognition [2, 3, 4, 5], personalized handwriting recognition systems [6], automatic forensic document examination [7], classification of ancient manuscripts [8] and smart meeting rooms [9]. That is the reason why many efforts have been made in order to improve writer recognition methods.

Up to now, researchers in the field of Text-Independent Writer Recognition have mainly focused on the statistical approach. This has led to the specification and extraction of statistical features such as slant distribution, entropy, and edge-hinge distribution. We found that the edge-hinge distribution feature outperforms all other statistical features [2]. Therefore, the aim of this paper is to compare our improved run-lengths features with edge hinge features.

The remaining of the paper is organized as follows: in the first section, we give a brief overview on some significant recent contributions to Arabic writer recognition. In the next part, we introduce the database used in our study, followed by the description of our proposed approach. The following section presents the experimental results and their analysis. Finally, we give a conclusion with some future research directions.

2   Arabic Writer Recognition: A Survey

Writer recognition from Arabic handwritten documents has not been addressed as extensively as writer recognition from Latin or Chinese handwritten documents until the last few years. The first study dates back to 2005 when Al-Zoubeidy et al [10] proposed the use of multichannel Gabor filtering and gray-scale co-occurrence matrices to characterize the writing style of writers. Gazzah et al [11] combined local and global features. Global features were extracted with 2D DWT using lifting scheme but the local features describe the morphological variations of writing (lines height, ascenders slant and diacritical dots features).

Al-Dmour et al [12] presented a feature extraction technique based on hybrid spectral-statistical measures (SSMs) of texture. Bulacu et al [2] proposed an approach based on the combination of textural with allographic features. Joint directional probability distributions and grapheme-emission distributions are extracted independently of the textual content of the written samples. The authors conducted an analysis of the combination of textural and allographic features and showed that the combination of these features improves the performances.

Abdi et al [13] proposed a method based on the combination and the cooperation of six feature vectors computed from the minimum perimeter polygon (MPP) contours of Arabic words. These feature vectors are in the form of probability distribution functions (PDFs), and are based on the length, direction, angle and curvature measurements. In [14] the authors calculate the fractal dimensions for images by using the "Box-counting" method, and calculate the multi-fractal dimensions of images by using the method of DLA (Diffusion Limited Aggregates).

Awaida et al [15] addresses writer identification of Arabic handwritten digits. A combination of Gradient, curvature, density, horizontal and vertical run lengths, stroke, and concavity features is used for characterization of the writing samples.

Al-Ma'adeed et al [16] evaluated the performance of edge-based directional probability distributions as features and moment invariants and words' measurements such as area, length, height, length from baseline to upper edge and length from baseline to lower edge in Writer identification.

Chen et al [17] proposed a method for detecting and removing ruling lines from the handwritten documents and tested its utility for Arabic Writer identification through series of experiments. Their preliminary results show that, under realistic assumptions where ruling lines are expected to have different properties across the collection, e.g., thickness, spacing, etc., removing them significantly improves identification performances.

3   Feature Extraction

In our work, two texture analysis methods are implemented and used for characterizing Arabic handwritings, these methods are : Run-Length distribution [18] and edge-hinge distribution [19].

3.1   Run-Length Features

To characterize the writing style of different writers involved in the evaluation of our writer recognition methods, we compute the probability distribution of run-lengths features, which are determined on a binary image taking into consideration the black pixels corresponding to the ink trace and the white pixels corresponding to the background.

There are four scanning methods: horizontal, vertical, left-diagonal and right-diagonal. We calculate the runs-lengths features using the grey level run-length matrices and the histogram of run-lengths is normalized and interpreted as a probability distribution function (PDF). The method considers horizontal, vertical, left-diagonal and right-diagonal white run-lengths as well as horizontal, vertical, left-diagonal and right-diagonal black run-lengths extracted from the image of the handwritten document.

The run-lengths features we propose to use here give information on the average width of the letters, the density of writing, the structure of the letters, the average size of the letters, the ink width, the characters position, the regions enclosed inside the letters and also the empty spaces between letters and words, the regularity and irregularity of handwriting and finally the slope in handwriting.

We have used the set of features proposed here in the ICDAR'2011 Writer Identification Contest [20], we have also used a part of these features in the ICDAR'2011 Arabic Writer Identification Contest [21] and in the ICDAR'2011 Music Scores Competition: Staff Removal and Writer Identification [22]. We have obtained interesting results in these competitions.

3.2   Edge-Hinge Features

Edge-hinge distribution is a feature that characterizes the changes in direction of a writing stroke in handwritten text [19]. The edge-hinge distribution is extracted by

means of a window that is slid over an edge-detected binary handwriting image. Whenever the central pixel of the window is on, the two edge fragments (i.e. connected sequences of pixels) emerging from this central pixel are considered. Their directions are measured and stored as pairs. A joint probability distribution $P(\varphi1, \varphi2)$ is obtained from a large sample of such pairs. An example of an angle pair is shown in Figure 1.
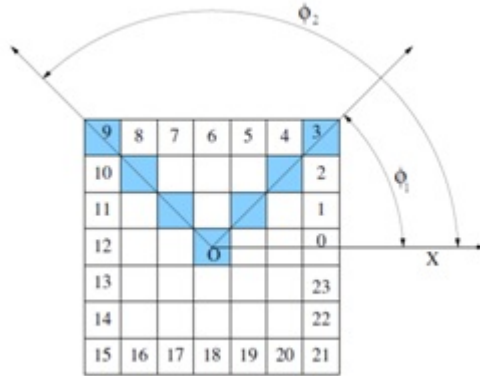

Fig. 1. Example of an edge-hinge distribution (image reproduced from [19])

4   Writer Recognition

Once the handwriting samples have been represented by their respective features, we need to compute the distances between respective features to define a (dis)similarity between two handwriting samples. We tested three distance measures including: Euclidean distance, Chi-square distance and Manhattan distance. In our experimentations, Manhattan distance performed the best.

In writer identification task, the efficiency of the considered features has been evaluated using nearest-neighbor classification [2] in a leave-one-out strategy. Explicitly, one document (a query document) is chosen and extracted from the total of 1375 documents (note that the experimental dataset contains 5 documents written by each of 275 writers), then the distances between the features vector of the chosen document and the features vectors of all of the remaining 1374 documents are computed. For a query document, we don't only find the Top-1 but a longer list up to a given rank (Top-10) thus increasing the chance of finding the correct writer in the retrieved list.

For writer verification, we compute the distance between two given documents and consider them as being written by the same writer if the distance falls within a predefined decision threshold. Beyond the threshold value, we consider that the documents are written by different writers. By varying the acceptance threshold, the ROC curves are computed and the verification performance is quantified by the Equal Error Rate (EER): the point on the curve where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). The lower is EER value, the higher is the accuracy of the system.

5   Experimental Study

The experimental study was carried out on the writing samples from the IFN/ENIT database [23] which is the unique Arabic handwriting publicly available database. It consists of forms with handwritten Arabic town/village names collected from 411 subjects (binary images at 300 dpi resolution). Most writers filled in 5 forms. This database was designed for training and testing recognition systems for handwritten words and was used for the ICDAR 2005 Arabic OCR competition [24].

The IFN/ENIT database was used also in [2, 3, 4, 5, 13] for writer identification and verification because the writer information was recorded. We have extracted the handwriting from the scanned forms. The text content is variable and the samples contain a limited amount of handwriting: only 12 names and 12 zip codes of Tunisian towns/villages. In our writer identification and verification experiments, we used the data concerning 275 writers with 5 samples per writer.

Table 1. Overview of proposed features and their dimensions

Feature	Description	Dimension
f1	Horizontal run-lengths on white pixels	120
f2	Left-diagonal run-lengths on white pixels	120
f3	Vertical run-lengths on white pixels	120
f4	Right-diagonal run-lengths on white pixels	120
f5	Horizontal run-lengths on black pixels	264
f6	Left-diagonal run-lengths on black pixels	264
f7	Vertical run-lengths on black pixels	264
f8	Right-diagonal run-lengths on black pixels	264
f9	Edge-hinge with fragment of 5 pixels	1024
f10	Edge-hinge with fragment of 6 pixels	1600
f11	Edge-hinge with fragment of 7 pixels	2304
f12	Edge-hinge with fragment of 8 pixels	3136

To evaluate the proposed approach, we have conducted two types of experiments: the first one is designed to evaluate the result we can reach by using individually each studied feature vector. Whereas the second type aims at testing the result we can reach by combining the studied feature vectors. For writer identification task, we report the Top 1, Top 5 and Top 10 identification rates while for writer verification task, we present the Equal-Error-Rate (EER).

For each feature, Table 1 summarizes the corresponding number, the description and the dimension, whereas Table 2 presents the performance of the individual features detailed in the above sections. Although the feature performances vary signif-icantly, it can be noticed that the edge-hinge features (f9-f12) outperform the run-lengths features (f1-f8), with f12 (Edge-hinge with fragment of 8 pixels) achieving the best results both on identification and verification tasks.

Table 2. Writer recognition performance on individual features

Feature	Top 1	Top 5	Top 10	EER
f1	14,27%	36,51%	50,91%	17,58%
f2	16,07%	41,31%	56,44%	15,02%
f3	8,94%	26,25%	38,47%	21,29%
f4	17,89%	42,40%	56,00%	16,07%
f5	28,65%	56,51%	69,16%	14,20%
f6	28,65%	54,62%	67,27%	13,83%
f7	29,96%	53,16%	65,02%	15,62%
f8	30,69%	54,98%	65,96%	15,14%
f9	83,56%	95,49%	97,45%	6,58%
f10	84,36%	95,34%	97,45%	7,08%
f11	87,49%	97,02%	97,82%	6,30%
f12	89,16%	97,45%	98,84%	5,49%

Table 3. Writer recognition performance on features combination

Features combinations	Top 1	Top 5	Top 10	EER
f1, f2, f3, f4	47,20%	76,14%	86,54%	10,56%
f5, f6, f7, f8	75,42%	90,25%	93,82%	9,56%
f1, f2, f3, f4, f5, f6, f7, f8	88,07%	96,87%	98,54%	5,80%
f1, f2, f3, f4, f5, f6, f7, f8, f12	93,53%	98,47%	99,13%	4,78%

Table 3 summarizes some of the combinations we have tested. For writer identification, the highest rate we have reached stands at 93.53% in Top 1, 98.47% in Top 5 and 99.13% in Top 10 when combining run-lengths on white and black pixels with edge hinge with fragment of 8 pixels (f1-f8, f12). For the verification task, we achieve an EER of 4.78% when combining run-lengths on white and black pixels with edge hinge with fragment of 8 pixels (f1-f8, f12). The ROC curves for some of the feature combinations have been illustrated in figure 2.

When comparing the recognition performance across the two types of features, it can be seen that the identification and verification results are much poor when using run-lengths with individual features but that is comparable with the edge hinge features when we combine all the run-lengths features. Since the IFN/ENIT database has been widely used in evaluating writer identification and verification tasks, it would be interesting to present a comparative overview of the proposed methods.

Table 4 summarizes the performance of the most recent studies on writer identification and verification on this dataset. Bulacu & al [2] currently hold the best performance results with 88% in Top 1 and 99% in Top 10 on 350 writers in identification task and EER of 5.8% in verification task. We have achieved an identification rate of 88.07% in Top 1, 96.87% in Top 5 and 98.54% in Top 10 by using the run-lengths features and we have improved the results by combining the run-lengths features with edge hinge features to achieve an identification rate of 93.53% in Top 1, 98.47% in Top 5 and 99.13% in Top 10 and an EER of 4.78%.
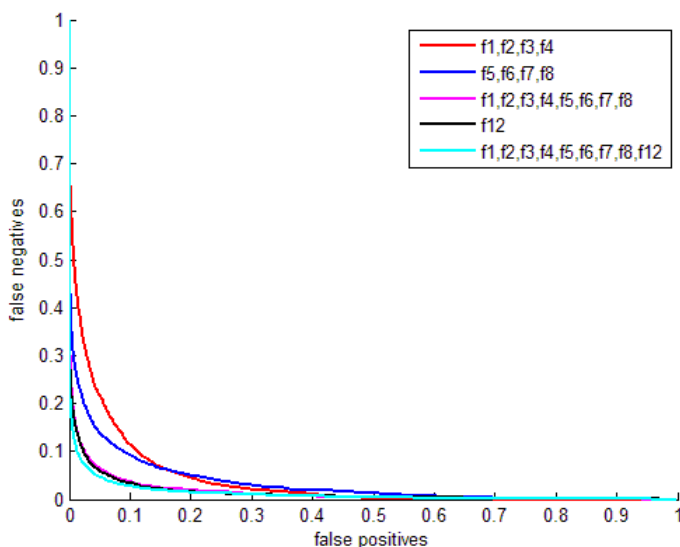
Fig. 2. ROC curves for some of the feature combinations

Table 4. Comparison of writer recognition methods

Reference	Writers	Top 1	Top 5	Top 10	ERR
Abdi & al [13]	82	90.20%	96.30%	97.50-%	-
Bulacu & al [2]	350	88.00%	-	99.00%	5.80%
Our method	275	93,53%	98,47%	99,13%	4,78%

6 Conclusion and Future Work

We have proposed here a new writer recognition method based on Arabic handwriting. The strength of this method is demonstrated experimentally by the classification of 1375 Arabic handwriting images from 275 different writers. Comparisons of improved run-lengths features with the edge hinge features demonstrate that the run-lengths features possess good discriminatory information and that a good method of extracting such information is the key to success of the classification.

The method that has been proposed here is mainly text-independent. In our future work, text-dependent writer identification will be considered and can include signature verification methods. A comparison between the two approaches will then be conducted. Currently, our work is based on the extraction of global features, but further work will focus on the use of local features. An integrated system will be considered by combining both local and global features to produce more reliable classification accuracy.

References

1. Schlapbach, A.: Writer Identification and Verification. PhD Thesis, Bern University (2007)
2. Bulacu, M., Schomaker, L., Brink, A.: Text-Independent Writer Identification and Verification on Off-Line Arabic Handwriting. In: 9th International Conference on Document Analysis and Recognition, Brazil, vol. 2, pp. 769–773 (2007)
3. Djeddi, C., Souici-Meslati, L.: Une approche locale en mode indépendant du texte pour l'identification de scripteurs: Application à l'écriture arabe. In: Colloque International Francophone sur le Document et l'Ecrit, Rouen, France, pp. 151–156 (2008)
4. Djeddi, C., Souici-Meslati, L.: A texture based approach for Arabic Writer Identification and Verification. In: IEEE International Conference on Machine and Web Intelligence, Algiers, Algeria, pp. 88–93 (2010)
5. Djeddi, C., Souici-Meslati, L.: Artificial Immune Recognition System for Arabic Writer Identification. In: 4th IEEE International Symposium on Innovation in Information and Communication Technology, Amman, Jordan, pp. 159–165 (2011)
6. Nosary, A., Heutte, L., Paquet, T.: Unsupervised writer adaption applied to handwritten text recognition. Pattern Recognition 37, 385–388 (2004)
7. Van Erp, M., Vuurpijl, L., Franke, K., Schomaker, L.: The WANDA measurement tool for forensic document examination. Journal of Forensic Document Examination 16, 103–118 (2005)
8. Siddiqi, I., Cloppet, F., Vincent, N.: Contour Based Features for the Classification of Ancient Manuscripts. In: 14th Conference of the International Graphonomics Society, Dijon, France (2009)
9. Liwicki, M., Schlapbach, A., Bunke, H., Bengio, S., Mariéthoz, J., Richiardi, J.: Writer Identification for Smart Meeting Room Systems. IDIAP research report IDIAP-RR 05-70 (2005)
10. Al-Zoubeidy, L.M., Al-Najar, H.F.: Arabic writer identification for handwriting images. In: International Arab Conference on Information Technology, Amman, Jordan, pp. 111–117 (2005)
11. Gazzah, S., Ben Amara, N.E.: Arabic Handwriting Texture Analysis for Writer Identification using the DWT-lifting Scheme. In: 9th International Conference on Document Analysis and Recognition, vol. 2, pp. 1133–1137 (2007)
12. Al-Dmour, A., Abu Zitar, R.: Arabic Writer Identification based on Hybrid Spectral-Statistical Measures. Journal of Experimental and Theoretical Artificial Intelligence 19, 307–332 (2007)
13. Abdi, M.N., Khemakhem, M., Ben-Abdallah, H.: Off-Line Text-Independent Arabic Writer Identification using Contour-Based Features. International Journal of Signal and Image Processing 1, 4–11 (2010)
14. Chaabouni, A., Boubaker, H., Kherallah, M., Alimi, A.M., El Abed, H.: Fractal and Multifractal for Arabic Offline Writer Identification. In: International Conference on Pattern Recognition, Istanbul, Turkey, pp. 1051–4651 (2010)
15. Awaida, S.M., Mahmoud, S.A.: Writer Identification of Arabic Handwritten Digits. In: 1st International Workshop on Frontiers in Arabic Handwriting Recognition, Istanbul, Turkey (2010)
16. Al-Ma'adeed, S., Mohammed, E., Al Kassis, D., Al-Muslih, F.: Writer Identification using Edge-Based Directional Probability Distribution Features for Arabic Words. In: IEEE/ACS International Conference on Computer Systems and Applications, pp. 582–590 (2008)

17. Chen, J., Lopresti, D., Kavallieratou, E.: The Impact of Ruling Lines on Writer Identification. In: 12th International Conference on Frontiers in Handwriting Recognition, Kolkata, India, pp. 439–444 (2010)
18. Tang, X.: Texture Information in Run-Length Matrices. IEEE Transactions on Image Processing 7(11), 1602–1609 (1998)
19. Bulacu, M.: Statistical Pattern Recognition for Automatic Writer Identification and Verification. PhD thesis, University of Groningen (2007)
20. Louloudis, G., Stamatopoulos, N., Gatos, B.: ICDAR 2011 - Writer Identification Contest. In: 11th International Conference on Document Analysis and Recognition, Beijing, China, pp. 1475–1479 (2011)
21. Hassaine, A., Al-Maadeed, S., Alja'am, J.M., Jaoua, A., Bouridane, A.: The ICDAR 2011 Arabic Writer Identification Contest. In: 11th International Conference on Document Analysis and Recognition, China, pp. 1470–1474 (2011)
22. Fornés, A., Dutta, A., Gordo, A., Llados, J.: The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification. In: 11th International Conference on Document Analysis and Recognition, Beijing, China, pp. 1511–1415 (2011)
23. Pechwitz, M., Maddouri, S., Margner, V., Ellouze, N., Amiri, H.: IFN/ENIT-database of handwritten arabic words. In: Colloque International Francophone sur le Document et l'Ecrit, pp. 129–136 (2002)
24. Margner, V., Pechwitz, M., El Abed, H.: ICDAR 2005 arabic handwriting recognition competition. In: 8th International Conference on Document Analysis and Recognition, pp. 70–74 (2005)