2011 International Conference on Document Analysis and Recognition

Multi-Fractal Modeling for On-line Text-Independent Writer Identification

Aymen Chaabouni, Houcine Boubaker
Monji Kherallah and Adel M. Alimi
REGIM: REsearch Group on Intelligent Machines, University of Sfax,
National School of Engineers BP 1173, Sfax, 3038, Tunisia
{*ayman.chaabouni, houcine-boubaker, monji.kherallah, adel.alimi*}*@ieee.org*

Haikal El Abed
Technische Universität Braunschweig,
Institute for Communications Technology
(IfN),Braunschweig, Germany
elabed@tu-bs.de

***Abstract*—The aim of this paper is to address the task of writer Identification of on-line handwriting. A new method for analytical on-line writer identification is proposed. However, although it is possible to measure the degree of handwriting irregularity thanks to the fractal dimension, the fractal analysis with a single exponent is not enough sufficient to characterize handwriting styles variation; instead, a continuous spectrum of exponents is necessary. In this purpose Multi-Fractal analysis was used to characterize styles of writing of writers. The main objective of this study is to explore the utility of this novel statistical tool for the purpose of distinguishing styles of on-line writings. Furthermore, a new method to estimate Multi-Fractal dimensions for on-line handwriting is presented and a procedure to find the most distinctive graphemes is elaborated. To evaluate our method, we have used the writings of 100 writers from the ADAB database. Our experimental results demonstrate the effectiveness of our proposed method and show a large capability of Multi-fractal features to characterize on-line handwriting styles.**

***Keywords*-Multi-Fractal Modeling; On-line; Text-Independent; Distinctive Graphemes; Arabic/Persian Writer Identification**

I. Introduction

The emergence of new technologies in the field of data entry and data collection have made the appearance of devices capable of producing on-line handwriting documents(tablet PC, electronic pen, PDA, etc.). The very rapid growth of these documents raises the question of access to information contained in these documents. One of the important information is to find the identity of the author of a given document.

Writer identification is very useful in daily life. It arises in several tasks, especially whene we connect to the computer network, when we execute bank transactions in financial activities, in the legal world where it is fundamental to recognize the identity of a person as in the case of threat or ransom letters, bill of sales, wills, etc.

During the last decade, several approaches have been made in the writer identification field. These approaches are generally categorized as off-line, where only a scanned image of the handwriting is available, and On-line, where temporal and spatial information about the writing is available [1],[2]. Among these approaches, we can cite the system of Said et al [3], who presents an off-line approach based on gabor filter and the calculation of co-occurrence matrix. Bulacu et al [4] used the edge-based directional probability distributions as features for writer identification. Bensefia [5], presented an approach based on the invariants of writer. This approach is based on the comparison of the respective graphemes of documents, by a measure of similarity. Schlapbach et al [6], presented an on-line system of writer identification based on Gaussian mixture models. In other works, Chapran and Fairhurst [7] presented a new method for dynamic writer identification which uses the relation between static and dynamic information.

Recently, works which treat the Arabic handwriting were proposed. A study was made by Al Zoubeidy et al [8] has adapted the approach proposed by Said et al [3]. The features are extracted from the image of writing using a gabor filter and calculation of co-occurrence matrix . Another work was made by Bulacu et al [9], where probability distribution functions are extracted.

Over the past decade fractal analysis [10] has been applied on handwriting analysis and writer identification. An approach based on the study of the fractal behavior of handwriting has been proposed by Vincent [11] that demonstrates that is possible to characterize handwriting irregularity with the fractal analysis. Hertel and Bunke [12] have shown that The best performance among set of features is achieved by the fractal analysis in a writer identification system. Also, in the study [13], is demonstrated that fractals are powerful statistical tool for off-line writer identification and is demonstrated also that multi-fractal is more efficient than the simple fractal.

This paper proposes a new method for Arabic/Persian on-line Text-Independent writer identification. Our method is based on Multi-Fractal modeling. The suitability of using multi-fractal features for on-line handwriting is supported by previous use in off-line writer identification [13] and in other domains [14],[15],[16]. In this way, this study describes how multi-fractal analysis can be applied to characterize on-line handwriting styles. The main contributions of our study are: 1) It is the first attempt to apply Multi-fractal technique to characterize on-line handwriting styles, 2) A development of a new method to estimate Multi-Fractal dimensions for

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.131

on-line handwriting and 3) A procedure to find the most distinctive graphemes is elaborated.

This paper is organized as follows: in the first section we present multi-fractal analysis. In the next part, the framework for writer identification is presented. The following section presents experiments and results and finally we give a conclusion with some future work.

II. Multi-Fractal Analysis

In this section, we present the multi-fractal technique applied to the images and its applicability on the on-line handwriting.

A. Multi-Fractal Dimensions Dq

In contrast to simple fractal, multi-fractal is characterized by a continuous spectrum of exponents, rather than a single fractal dimension. this hierarchy of exponents allows to describe the fractal properties at different scales.
Given a binary image with number of black pixels M_0 and the size L, covered grid boxes of size l, the Multi-Fractal dimensions Dq for this image is defined as follows [14],[15]

$$\sum_i \left(\frac{M_i}{M_0}\right)^q \approx \left(\frac{l}{L}\right)^{(q-1)Dq} \tag{1}$$

Where M_i is the number of pixels in the ith box, and q is a variable which allows to distinguish fractals properties at different scales. A large difference between the fractal objects and the multifractal objects, is that for the mono-fractals, Dq is the same for all q : $D_{-\infty} = ... = D_0 = D_1 = D_2 = ... = D_{+\infty}$, On the contrary, if the multi-fractal dimensions decreases when q is increased: $D_{-\infty} > ... > D_0 > D_1 > D_2 > ... > D_{+\infty}$. For $q = 0$, D_0 corresponds to the fractal dimension,for $q = 1$, D_1 corresponds to the $information$ or $entropy$ dimension, and for $q = 2$, D_2 corresponds to the $correlation$ dimension. It turns out that the direct application of (1) in practice is hindered by the fact that for $q < 0$, the boxes that contain a small number of pixels give anomalously large contribution to the sum on the left hand side of (1) [16]. To solve this problem, a solution has been proposed in [14]. This solution is based on the application of $Generalized\ sand\ box\ method$ that is used to demonstrate the multifractality of the DLA (Diffusion Limited Aggregates). This procedure consists to choose randomly N pixels belonging to the structure, and counting for every pixel i the number of pixels M_i, inside boxes of linear dimension R, centered on the selected pixel. The left-hand side of the Equation (1) can be interpreted as the average of the quantity $\left(\frac{M_i}{M_0}\right)^{(q-1)}$ According to the probability distribution $\left(\frac{M_i}{M_0}\right)$, when the centers of the boxes are chosen randomly, the averaging is made during this distribution, and consequently, the Equation (1) becomes:

$$\langle\left(\frac{M(R)}{M_0}\right)^{(q-1)}\rangle \approx \left(\frac{R}{L}\right)^{(q-1)Dq} \tag{2}$$

where the $\langle ... \rangle$ denotes the average over the centers.

B. Application to the On-line Handwriting

In the on-line approaches, we can exploit more information on the writing, than off-line approaches. We can exploit the dynamics of writing, the speed of movement of the pen, the accelerations, the exercised pressure, and the temporal order of the writing, which is impossible to recover in the off-line approaches. For this purpose, we have adapted the method of DLA presented in previous sub-section to be applicable on on-line handwriting.

To calculate the multi-fractal dimensions Dq versus q, all points of the writing are taken, one after the other, by following the temporal order of the handwriting, until reaching the last point. For every box of radius R_i=(R_1, R_2,...,R_{max}) centered on current point, we calculate the number of points inside this box, by counting only the points which are in the temporal order before the current point. Thereafter, the counts were used to calculate $\log\langle(M_i(R)/M_0)^{q-1}\rangle/(q-1)$ versus $\log(\frac{R}{L})$ for each value of q, where the $\langle ... \rangle$ denotes the average over the centers. Dq values correspond to the slopes of the straight lines, obtained by least squares fitting.

$$D_q = \begin{cases} \frac{1}{q-1}\frac{\log\left[\left(\frac{M(R)}{M_0}\right)^{q-1}\right]}{\log\left[\frac{R}{L}\right]}, & \text{if } q \neq 1 \\ (D_{q+\varepsilon} + D_{q-\varepsilon})/2, & \text{if } q = 1 \end{cases} \tag{3}$$

Figure 1 shows the application of multi-fractal for on-line handwriting.

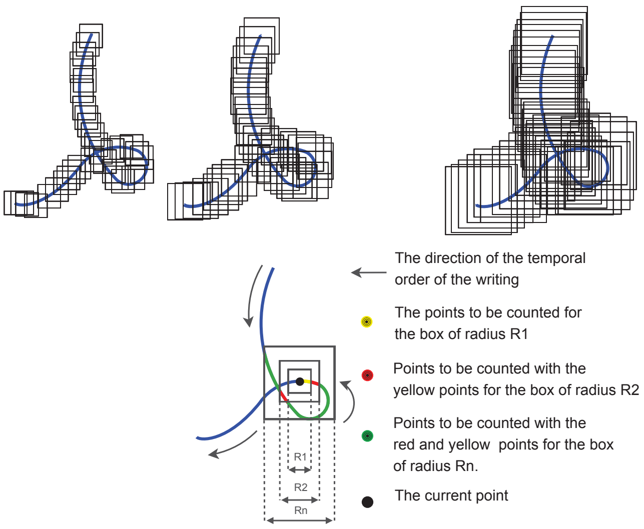
The direction of the temporal order of the writing
The points to be counted for the box of radius R1
Points to be counted with the yellow points for the box of radius R2
Points to be counted with the red and yellow points for the box of radius Rn.
The current point
R1
R2
Rn

Figure 1. Multi-Fractal Dimensions for On-line Handwriting

III. Framework for Writer Identification

A. Data Preprocessing and Graphemes Segmentation

The preprocessing stage aims to eliminate the noise and to normalize the handwriting size. First, we use a Chebyshev second type low pass filter with a cutoff frequency of f_{cut}=12Hz to eliminate the noise introduced by temporal and spatial sampling. Then, the vertical dimension of the handwriting lines is adjusted to obtain a normalized size script.

The segmentation of the Arabic/Persian pseudo-words in graphemes is based on the detection of two types of topologically particular points: Bottom of the ligature valleys and Angular points [17],[18]. The set of the segmented graphemes is decomposable in 5 groups: graphemes in the beginning, the isolated graphemes, the diacritics, the graphemes in the middle and the graphemes in the end.

B. Multi-Fractal Features Extraction

The stage of Multi-Fractal features extraction is based on the application of multi-fractal modeling on the on-line writing presented in previous section. It consists to extract the multi-fractal dimensions of graphemes resulting from the graphemes segmentation process. Afterward, we collect 5 small databases of Multi-Fractal dimensions of graphemes of all writers. These databases relating to: graphemes in beginning, isolated graphemes, diacritics, graphemes in the middle and graphemes in the end (Figure 2).

C. Identification Measurement

For each one of the 5 groups of multi-fractal dimensions of graphemes, we apply the K-means clustering algorithm, with a determined number of sub-groups K defined as follows: $K_g = N_w \times N_pG_g$
where N_w is the number of writers and N_pG_g is the number of prototypes of graphemes in a given group g. The classification of different forms of graphemes according to their position in the Arabic/Persian words, gave the following statistics: graphemes in the beginning: 11 prototypes, isolated graphemes: 13 prototypes, diacritics: 5 prototypes, graphemes in the middle: 13 prototypes, graphemes in the end: 13 prototypes.

The process of calculation of score is similar to that of information retrieval(IR) that allows to determine which document is most relevant to a query. The query in the process of identifying the writer consists of several graphemes and the information sought is the author of these graphemes. For this, we have adapted the technique $TFIDF$ (Term Frequency-Inverse Document Frequency) [19] used in IR. The use of IR theory has been successfully applied in the works [20],[21]. The graphemes frequency (TF) allows to estimate the importance of the graphemes of a given writer i contents in a subgroup j. TF is defined by: $TF_{i,j} = \frac{N_{i,j}}{\sum N_{i,j}}$. Where $N_{i,j}$ is the number of graphemes of the writer i in subgroup j.

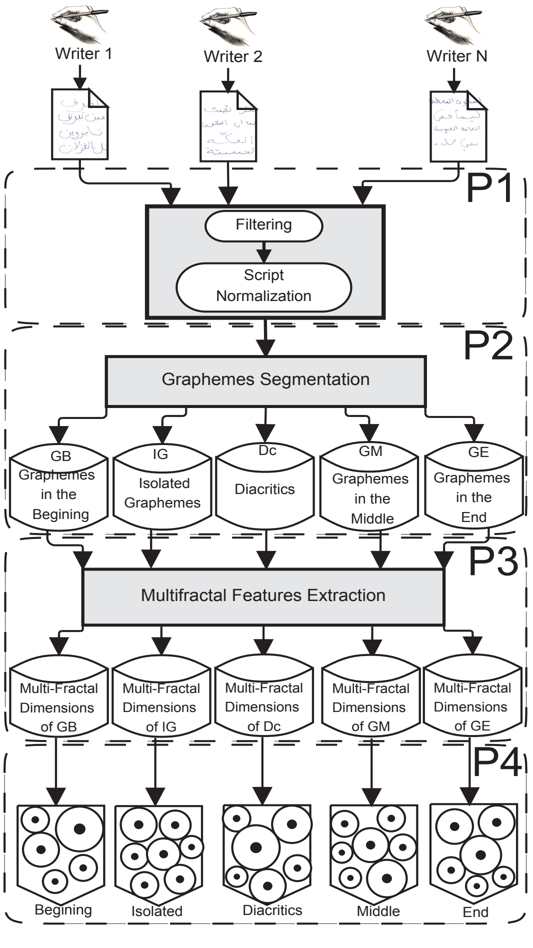
Writer 1
Writer 2
Writer N
P1
Filtering
Script Normalization
P2
Graphemes Segmentation
GB
Graphemes in the Begining
IG
Isolated Graphemes
Dc
Diacritics
GM
Graphemes in the Middle
GE
Graphemes in the End
P3
Multifractal Features Extraction
Multi-Fractal Dimensions of GB
Multi-Fractal Dimensions of IG
Multi-Fractal Dimensions of Dc
Multi-Fractal Dimensions of GM
Multi-Fractal Dimensions of GE
P4
Begining
Isolated
Diacritics
Middle
End

Figure 2. Training Process
***P4** represents the subgroups resulting from K-means clustering algorithm for each group*

Nevertheless, if the graphemes of a writer i are themselves very frequent in subgroups, IE they are present in many subgroups, they are somewhat discriminating and consequently they cannot discriminate the writers. That is why, it is necessary to increase the relevance of the graphemes of the writer i according to their rarity in subgroups, and this by the calculation of the inverse subgroup Frequency of (IDF). IDF is a measure designed to give a more important weight for the graphemes of a writer which are less frequent and form a subgroups. IDF is defined as: $IDF_i = \log(\frac{N_{SG}}{N_{SG_i}})$. Where N_{SG} is the total number of subgroups and N_{SG_i} is the number of subgroups in which they appeared the graphemes of the writer i. The weight increases proportionally with the frequency of graphemes of the writer i in the subgroup j ($TF_{i,j}$) and according to the frequence of these graphemes in all the subgroups (IDF_i). Finally the weight is obtained by multiplying both measures: $TF_{i,j} \times IDF_i$.

In the identification phase as demonstrated in Figure 3, after the execution of processes: P1, P2 and P3, we obtain 5 small databases of multi-fractal dimensions of graphemes. Each grapheme is affected to the subgroup that has a minimum euclidean distance with its vector of multi-fractal

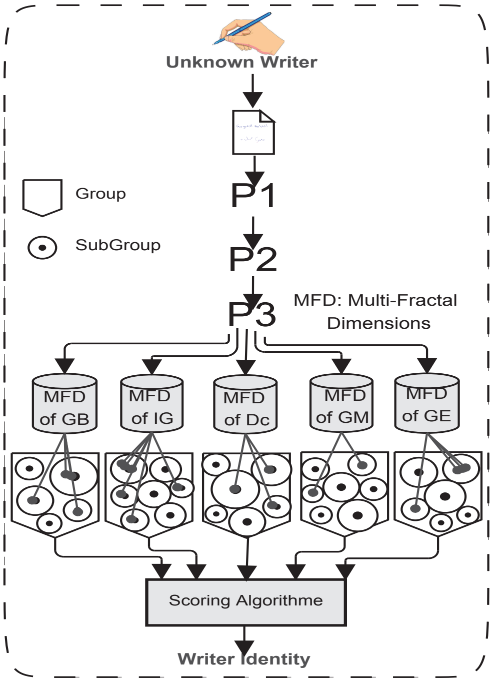

Figure 3. Identification Process

dimensions. Once all graphemes are assigned to subgroups (Figure 3), we apply the algorithm scoring that is based on the following formula:

$$score(i,g) = \sum_{j=1}^{K_g} N_G(i,j,g) \times TF_{i,j,g} \times IDF_{i,g} \qquad (4)$$

Where $N_G(i,j,g)$ is the number of graphemes of writer i affected in the subgroup j of group g. Equation (4) defines the score of each writer for each group of graphemes (isolated, beginning, middle, diacritics, end). The final score is defined as follows:

$$score(i) = \frac{1}{5}\sum_{g=1}^{5} score(i,g) \qquad (5)$$

D. Distinctive Graphemes Detection

The procedure for detecting the distinctive graphemes requires two training levels. The first level is similar to that shown in Figure 2 and the second level, consist to define a weight α, which allows to promote the most distinctive graphemes among the 5 groups. That is why, we take again for each writer, the same samples used for the training of the first level, and consider them as test samples (Figure 3) to see the groups of the most distinctive graphemes of each writer i. The weight α is based on Equation (4) and is defined as follows:

$$\alpha_{i,g} = \frac{score(i,g)}{\sum_{g=1}^{5} score(i,g)} \qquad (6)$$

The final score becomes :

$$score(i) = \frac{1}{5}\sum_{g=1}^{5}(1+\alpha_{i,g}) \times score(i,g) \qquad (7)$$

IV. Experiments and Results

We have used the ADAB database [22] for our experiments, we have selected 100 writers, each one has wrote about about 1 page composed about 120 words, where two thirds of words have been used for the training phase, and the rest has been used for the identification. All experiments are performed on four stages, in a way that each writer wrote 10, 20, 30, and 40 words. After the stage of graphemes segmentation and features extraction, we obtain a database composed of 5 groups of multi-fractal dimensions relating to beginning graphemes, isolated graphemes, diacritics, graphemes in the middle and ending graphemes. The calculation of the multi-fractal dimensions is made for $-10 < q < 10$ to obtain 21 features for each grapheme. For the case where $q = 1$, D_q=$(D_{q+\varepsilon} + D_{q-\varepsilon})/2$, with $\varepsilon = 0.001$. The box sizes R are defined by: $R = n \times \Delta/50$, where Δ =(length of grapheme + width of grapheme)/2 and n is ranged from 2 to 32 in steps of 2.

Table I
Identification rate

Number of Words	Identification Rate	
	Top1	Top5
10	82.5%	97.5%
20	86.2%	97.5%
30	88.4%	100%
40	91.6%	100%

These experiments have shown that the results increase proportionally with the number of written words. Table I shows the evolution of results in proportion to the number of written words. the result was improved from 82.5% when the writers have written only 10 words to 91.6% when they have written 40 words.

Table II
Effect of the factor α on the results.

Number of Words	Identification Rate	
	Top1	Top5
10	87.8%	100%
20	90.1%	100%
30	92.2%	100%
40	95.5%	100%

Table II shows the improvement of the results through the factor α which means a statistical measure that favors the most distinctive graphemes for each writer. By introducing of this factor in the identification process, the result is improved from 82.5% to 87.8% when the writers have written 10 words and from 91.6% to 95.5% when they have written 40 words. The justification for the choice of (1 + α) in Equation (7) is explained by the fact that our orientation to identify the writer, was to favor the most distinctive graphemes, not to exclude the non-distinctive graphemes.

V. Conclusions and Futur Works

We presented in this paper a novel approach for on-line text-independent writer identification. The first system for on-line Arabic/Persian writer identification was developed. Our system is based on Multi-fractal technique. Previous researches in many domains especially in the field of image processing, have supported reliability and validity of this statistical tool. In this way, this study describes how multi-fractal features can be applied to characterize on-line handwriting styles. The best identification rate was 91.6% when writers wrote 40 words. In order to enhance the results, we have defined a weight α that promotes the groups of the most distinctive graphemes. This statistical measure has allowed to improve the results to 95.5%. The results of this study suggest that Multi-fractal modeling is a useful, yet promising technique for differentiating individuals based on their on-line writings.

However, our study is to be continued, we plan to work on other large-scale databases. We also intend to combine multi-fractal features with other features of literature. Also, it's necessary to note that multi-fractals are not language-dependent features, their applicability on other scripts will be an interest point in future research.

References

[1] M. Kherallah, L. Hadded, A. Mitiche, A. M. Alimi. On-Line Recognition Of Handwritten Digits Based On Trajectory And Velocity Modelling, *Pattern Recognition Letter* . Vol. 29. pp. 580-594. 2008.

[2] M. Kherallah, F. Bouri, A. M. Alimi. On-Line Arabic Handwriting Recognition System Based On Visual Encoding And Genetic Algorithm, *Engineering Applications of Artificial Intelligence*. Vol.22.pp.153-170 2009.

[3] HES. Said, G.S. Peake, T.N. Tan, K.D. Baker. Writer identification from non-uniformly skewed handwriting images, *Proceedings of the Ninth British Machine Vision Conference (BMVC 98), Southampton, England*. vol.2, pp. 478-487, 1998.

[4] M. Bulacu, L. Schomaker, L. Vuurpijl. Writer identification using edge-based directional features, *ICDAR'03, Edinburgh, Scotland*. pp 937-941, 2003.

[5] A. Bensefia, T. Paquet, L. Heutte. Grapheme based writer verification, *11th Conference of the International Graphonomics Society (IGS'2003), Scottsdale, Arizona*. pp. 274-277,2003.

[6] A. Schlapbach, M. Liwicki, H. Bunke. A writer identification system for on-line whiteboard data, *Pattern recognition*. pp.2381-2397, 2008.

[7] J. Chapran, M.C. Fairhurst, Biometric writer identification based on the interdependency between static and dynamic features of handwriting, *Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition*. pp. 505-510, 2006.

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

[9] M. Bulacu, L. Schmaker, A. Brink. Text-Independent Writer Identification and Verification on Offline Arabic Handwriting, *ICDAR 2007, Brazil*. Vol. 2, pp. 769-773, 2007.

[10] B. Mandelbrot. Les projets fractals, forme, hasard et dimensions. *Flammarion*. Paris, 1975.

[11] N. Vincent, H. Emptoz. A classification of writings based on fractals, *Fractal Reviews in the Natural and Applied Sciences. M. M. Novak. London: Chapman and Hall*. pp 320-331. 1995.

[12] C. Hertel, H. Bunke A set of novel features for writer identification, *Audio- and Video-Based Biometric Person Authentication*. pp. 679-687. 2003.

[13] A. Chaabouni, H. Boubaker, M. Kherallah, A. M. Alimi, H. El Abed. Fractal and Multi-fractal for Arabic Offline Writer Identification, *International Conference Pattern Recognition*. pp. 3793-3796, 2010.

[14] T. Vicsek. *Les projets fractals, forme, hasard et dimensions*, Fractal Growth Phenomena, 2nd ed, Singapore: World Scientific, 1993.

[15] T. Vicsek, F. Family, P. Meakin. Multifractal geometry of diffusion-limited aggregates, *Europhys. Lett.* vol. 12, pp. 217-222,1990.

[16] T. Stosic, B.D. Stosic. Multifractal analysis of human retinal vessels, *Medical Imaging, IEEE Transactions*. vol. 25, pp. 1101-1107, 2006

[17] H. Boubaker, A. El Baati, M. Kherallah, A. M. Alimi, H. El Abed. Online Arabic Handwriting Modeling System Based on the Graphemes Segmentation, *International Conference Pattern Recognition*. pp.2061-2064, 2010.

[18] H. Boubaker, A. Chaabouni, M. Kherallah, A. M. Alimi, H. El Abed. Fuzzy Segmentation and Graphemes Modeling for Online Arabic Handwriting Recognition, *International Conference on Frontiers in Handwriting Recognition* . pp.695-700, 2010.

[19] G. Salton, C. Buckley. Term-weighting approaches in automatic text retrieval, *Journal of Information Processing and Management*. pp.513-523, 1988.

[20] G. X. Tan, C. Viard-Gaudin, A. C. Kot, Automatic writer identification framework for online handwritten documents using character prototypes, *Pattern Recognition* vol. 42 pp. 3313-3323.

[21] A. Bensefia, T. Paquet, L. Heutte. A writer identification and verification system, *Pattern Recognition Letters* vol. 26, pp. 2080-2092, 2005.

[22] H. El Abed, M. Kherallah, V. Margner, A. M. Alimi ICDAR 2009 Arabic Online Handwriting Recognition Competition, *Proceedings of the 10th International Conference on Document Analysis and Recognition*. vol. 3, pp.1388-1392, 2009.