Online automatic Arabic handwritten signature and manuscript recognition

Ahmed Haddad/LABO RIADI
ENSI, university of Manouba
Tunis, TUNISIA
ahmed.haddad@ensi.rnu.tn

Amani Sfar/LABO RIADI
ISGS, university of Sousse
Tunis, TUNISIA
amanisfar.sfar@gmail.com

***Abstract*- The objective is to conceive and carry out an automatic verification system of the handwritten signatures while being based on the tools of recognitions of the manuscript writing based on the generic silhouettes, this system must allow a fast, systematic and effective checking of the signatures and would reduce the risks of counterfeit significantly. This system enters within the framework of software tools for the automatic treatment of the natural language. This modelling procedure will be based on a soft representation of words called "descriptor". We use the Arabic morpho-syntactic module of the linguistic platform NooJ for the automatic analysis of all the recognized forms.**

***Keywords*—TALN, handwritten, signature recognition, manuscript.**

I. INTRODUCTION

The signature of an individual, as a layout, results from a complex mechanism suitable for this one. It is thus supposed, that the signature of each individual is single, that it characterizes it. Obviously, several factors influence the layout of a signature: the position of the script writer, his mood, his physical shape. Consequently, certain variations (intra-individual) are present at the same person, one thus seeks to characterize a signature by taking into account these parameters. Variability will intra and interpersonal, the number and the types (cultural aspects) of signatures oblige with the settling, and with research, of methodologies to check and authenticate the signatures. Automatic generation of the dictionary

II. HANDWRITTEN SIGNATURE AND CHARACTER RECOGNITION

There exist currently several devices allowing the identification of an individual thanks to several characteristics biometric such as the size, the weight, the voice, the vascularization of the retina, the digital fingerprints, the signatures handwritten, etc an advantage of these characteristics is that they are specific to each individual and thus difficult to duplicate. In addition, acquisition, storage, the evolution in the time of certain parameters as their capacity discriminating within a broad population remain a constraint [14-16].

In addition, the handwritten signature of an individual represents a good compromise: while being relatively reliable, it is easy to acquire and it is socially well accepted like mode of identification. The signature is a means used since strong a long time to authenticate documents, for responsabiliser the individuals vis-a-vis engagements (contracts, etc). The signature is thus recognized like mode of validation associated with the identity with a person.

A. System of authentification

Our study is based on the use of a system of authentification conceived during preceding work [1]. This one requires two stages. At the time of the first, the phase of enrôlement (cf section 2.2), the person to be authenticated is recorded by providing some signatures so that the system creates its profile (together signatures of reference). The second phase is the authentification strictly speaking: a person - the beforehand recorded person or another who attacks it system subjects a signature which is authenticated or not. In all the cases, the acquisition of the signatures and the pretreatments are identical .

B. Acquisition of the signatures

The acquisition of the signatures is carried out on TabletPC, a graphics tablet or a PDA. In order to preserve the maximum of compatibility, only the coordinates of each point of the signature as well as posed and raised stylet are preserved. Then, a standardization is carried out on the signatures. This standardization consists of: a rotation according to the axis of inertia [2]; a homothety so that all the signatures have the same size while preserving their proportions; and finally in a translation to center the signature compared to the reference mark. Figures 1 and 2 illustrate the result of this transformation.

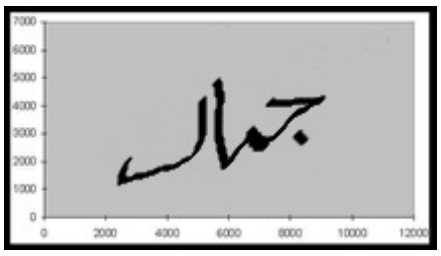

Fig. 1 . Original Signature with its axis of inertia.

978-1-5090-6778-7/17/$31.00 ©2017 IEEE

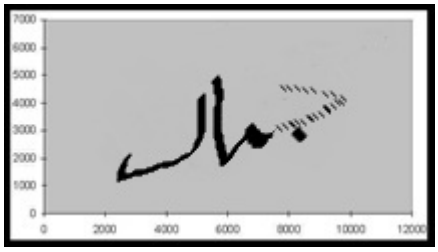

Fig. 2 . Character Signature

The last treatment carried out consists in preserving only certain significant points of the signature. The interest is multiple here. That makes it possible to reduce the size of the signature and thus to accelerate the treatments and that also allows to decrease storage space necessary for the profile. These two properties are particularly important for an application having to function within a real framework of use. Indeed, the CNIL prohibits the biometric databases. For each user, his biometric data must thus be stored on a personal smart card. Of more, to ensure the safety of the system, no information must be able to be recovered of this chart. This is why we store the profile of a user directly on a smart card. The authentication is also carried out on this chart. Thus the system is contestable with more difficulty, the counterpart being the use of limited resources system. A second advantage, and not of least, with the selection of points is that if those are quite selected, the performances of the system can be increased. Here, after comparative studies on several methods of selection of points, we have choose to preserve only the points minimal speed [3].

C. Enrôlement

This stage aims to create the profile of a user. After a phase of drive, necessary to the catch in hand of the peripheral of acquisition, the user provides 5 signatures to the system. In the event of problem during acquisition, the user can cancel a seizure and start again it. The system also carries out a checking to avoid any problem during the creation of the user profile. This checking consists in requiring 5 new signatures of the signatory. So among these 5 signatures of checking one among it is too dissimilar of the 5 signatures of references, the user must start again the procedure of enrolment. The dissimilarity between 2 signatures is evaluated by considering total time and the overall length of the signature, exactly in the same way that at the time of the first stage of our process of authentication.

Once the 5 acquired valid signatures, they undergo the treatments above (cf section 2.1) then are stored like signatures of reference on standby of a future authentication.

D. Authentification

At the time of this phase, a user wishing to authenticate sign on the peripheral of acquisition. The signature test thus obtained, after having undergone the described pretreatments, will be compared with the signatures of reference in order to determine if the signatory is well that which he claims to be. For that, the comparison is carried out in two stages, according to an approach Coarse To Fine [2]. The approach Coarse makes it possible to directly eliminate the signatures very different from the signatures from reference. These supposed attacks will thus be rejected without passing by the stage Fine which carries out a more precise comparison but also more expensive in computing times. This principle thus makes it possible to accelerate the treatments at the time of the authentication, the Fine stage being used only if necessary.

E. Fine stage

This second stage carries out a new comparison between the signature tested and the 5 signatures of reference, but this time by using a measurement of more precise distance. In our case, this distance is an alternative of DTW [3] adapted for the comparison of handwritten signatures in line [2]. Several metric can be associated with the DTW: the space distance, the temporal distance, the curvilinear distance [5]. Here, we use most powerful of the three, namely the space distance who corresponds to the Euclidean distance between two points put in correspondence and projected in the same reference mark.

So that a signature test is accepted, it is necessary that the distance between it and at least one of the signatures of reference is lower than the threshold µ. This total threshold is given at the time of a preliminary phase of training of the system. During this one, a base of validation is used to simulate the process of authentication and to obtain the performances of the system: TFA (Rate of Accepted Forgeries, or False Acceptance Misses - FAR) and TFR (Rate of False Rejection or False Rejection Misses - FRR, i.e the authentic rate of rejected signatures). Various threshold values are then tested in a quasi ehaustive way (by increment fixes) until obtaining the TEE (Equal Error Rate, or Equal Error Misses - EER) on this basis of validation.

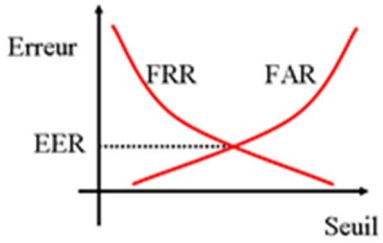

Fig. 3 - Definition of the TEE (or EER).

III. AUTOMATIC CREATION OF CLASSES OF SIGNATURES

Primarily resides in the use of the total threshold of decision µ which is single whatever the users of the system. This is why we wished to study the impact of the creation of classes of signatures on the performances of the system. In section 3.1 we describe the mechanism used to create the classes of signatures. Then, in section 3.2, we describe the changes that induces on the parameter setting of the system. Lastly, next section described new operation at the time of the authentication. That with only one class, i.e. all its signatures (of references and to authenticate) belong to the same class. It is the strong assumption on which this first study rests and we will see how the number of classes (K), the space of representation as well as the algorithm used to determine the classes influential on classification like on the performances of the system.

To obtain K classes, we use either K means [6], or Fuzzy C-means [7], on the whole of the signatures of reference which

978-1-5090-6778-7/17/$31.00 ©2017 IEEE

one lays out at the time of the phase of training. Then, we define the class of a signatory as that in which find the majority of its signatures of references (strict membership).

With regard to the space of representation, we initially chose to work on characteristics globales1. Among those, the total time of the signature as its overall length is recognized as being stable characteristics for a signatory. However, to go further, we also worked on a vaster whole of 12 characteristics, describing the shape and the dynamics of the signatures. These characteristics are:

− Overall length (overall length);

− graft Report/ratio displacements towards the left and the right-hand side (rapDepHor);

− graft Report/ratio displacements upwards and to the bottom (rapDepVer);

− Relationship between horizontal and vertical displacements (rapDepXY);

− Distance enters the first and the last point (distPremDer);

− average horitzontal Displacement (depHorMoy);

− average vertical Displacement (depVerMoy);

− Angle enters the horizontal one and the line uniting the first and the last point (angleSign);

− total Time (tpsTotal);

− Average acceleration (accMoy);

− vertical Mean velocity (vtsMoyVert);

− Mean velocity horitzontale (vtsMoyHor).

In order to determine the correlations between these characteristics we carried out an analysis in principal components (ACP). The corresponding circle of the correlations is given in figure 4. On this one one notes that the overall length and total time remain rather strongly correlated. One suspects whereas the algorithm of classification will be able to rest only on one to be able of representation of the space of description limited enough.

A. Optimization of the system classes of signatures

Once the classes of signatures obtained starting from the signatures of reference of the base of training, we use the signatures of the base of validation to define a threshold of decision μ K for each one of these classes K (k=1,… K). The principle is the same one as for the initial system except that one seeks people having access to a video of the person signing. In our work of experimentation, we are interested initially only in the performances on forgeries aléatoires2 and we thus use only 40x20 signatures. We are also conscious that this base has a limited size. However, it is about the one of the rare bases available and used like benchmark of reference. It also should be noted that we do not have knowledge a priori on the representativeness of this base as for the various styles of signatures and thus as for the number of classes. Indeed, the origin of the contributors is not known. Moreover, those did not sign with their usual signature but invented some for the occasion.

We carry out our tests in leave-one-out. 39 signatories are used for the training and the validation: the first five signatures of each one of 39 signatories are used as signatures of reference in order to simulate the phase of enrolment and to carry out the search for classes; the 15*39 remaining signatures are used as bases validation to seek the thresholds. Thus, each signature of the base of validation, is tested compared to each signatory of the base of training, i.e. it is compared with each corresponding signature of reference. At the time of the Fine stage, the threshold used to accept or reject the signature is the threshold μ K corresponding to the class of the signatures of reference. The TFA or the TFR of the class is then updated consequently. When all the base of validation was used, the threshold is modified then the TFA and TFR of each class are modified until obtaining the TEE on each class[10].

IV. MODELING OF THE DICTIONARY: FEATURES OF THE ARABIC LINGUISTICS

The Arabic alphabet is composed of 28 letters, whose shapes change according to the position in the word. The Arabic writing is semi-cursive in its two shapes, printed and handwritten. Indeed, an Arabic word is a related entity sequence entirely separated called pseudo-word. A word can be composed of one or several pseudo-words; it is due to the presence of characters that cannot be attached to their successor. Every pseudo-word is a sequence of linked letters; which gives the cursive aspect to this writing. Let's note that an isolated character can constitute a pseudo-word, as shown below:

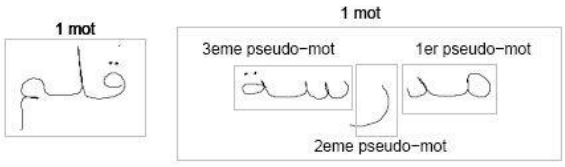

Fig. 4. Example of the composition an Arabic word

Arabic writing is rich in diacritical, and especially in points. There are 15 letters, among the 28 of the alphabet that include points. These points appear above (ض), below (ي), or in the middle of the character (ج).

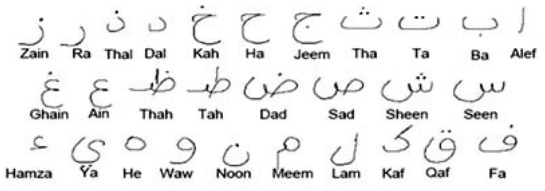

Fig. 5. isolated letters of the Arabic alphabet

The maximal number of diacritical that a letter can have is three points above the character or two below. These points permit to differentiate the pronunciation of Arabic letters. Therefore in a first level we propose a reduction based on the resemblance of

the general shape of characters in diacritical points, then we refine our modeling according to the generic silhouette notion in order to have a more developed reduction.

First reduction level: according to similar forms

As indicated previously, the Arabic alphabet is composed of 28 letters that have some common shapes. On this resemblance rests the idea of the reduction that consists in grouping letters that have similar diacritical features in classes of shapes, then one removes the diacritical points to finally have 18 distinct shapes. Hence, without the diacritical points, the letters " Ba", " Ta " and "Tha" have the same form.

Fig. 6 . Characters similar in form

If one assimilates letter "fa" to letter "qaf", which are only distinguished according to their position on or below the writing line, information which is not generally available unless one speaks of isolated characters or end of word characters, one has in this case 17 distinct shapes.

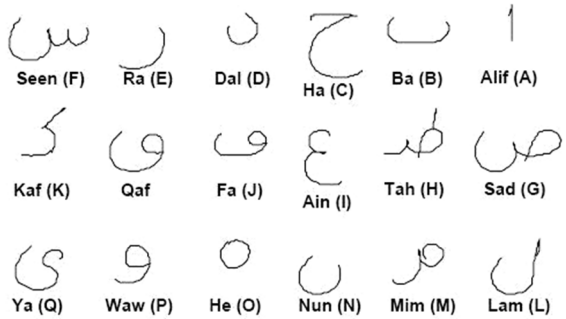

Fig. 7. the 18 forms of isolated Arabic characters

This first modeling permits to create a new reduced lexicon by simplifying the global shapes of characters. It is a stage that paves the way for a more developed reduction, the modeling by generic silhouette.

Static reduction of language: by generic silhouette

It is a specific modeling of the lexicon in view of aiming at a reduction in the time of a lexical post-treatment. The first stage consists in grouping characters in classes according to their size, their relative position in the word and the presence and the position of a vertical rod (information corresponding to the silhouette of characters), etc.

The silhouette of a character corresponds to the downward and relatively vertical fundamental features that compose it (features considered as insignificant are eliminated). The Arabic characters contain five downward features defined according to their position in the word:

– *medium* → does not exceed the body of the word,
– *upward/ hampe* → exceed upwardly,
– *downward / descendant / jambage* → exceed downwardly,
– *f-stroke* → exceed upwardly and dawnwardly.
– *curvature :* curves downwardly.

Label of feature	example
Medium	
Upward	
downward	
f-stroke	
curvature	

Fig. 8 . Fundamental features of the physical structure

The complete silhouette of a handwritten word is the concatenation of downward feature which composed. Then, it is necessary to index every word of the lexicon by an ideal describer which is its silhouette.

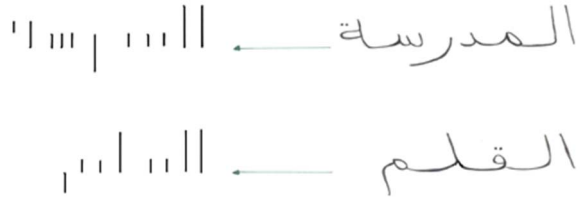

Fig. 9 . Example of a generic silhouette

During the lexical correction this information is taken into account and only substitutions of same class characters are accepted. The partition of the lexicon takes into account word size, its envelope (induced silhouette by silhouettes of characters) and the presence of character combinations.

As far as short words are concerned, only the information about the silhouette of characters is considered. As for the long ones, partitions are indexed by two characters and regroup all words containing these two characters. A word is placed therefore in all partitions designated by couples of characters that it contains, the number of words per partition is quite weak, but there are a lot of redundancies: on average a word is present eight times in the organized lexicon.

Modeling of proposed lexicon

Considering the existing means and the expected results, we opted for a static reduction of the lexicon, which is less time-consuming in calculation. The purpose is to generate the most reduced under-lexicons, while keeping a robust access criterion: the time of post-treatment decreases when the size of under-lexicons decreases, but the precision of the reduction also decreases. This reduction is characterized by:

1. The 18 distinct shapes of the Arabic alphabet, which are also called when designating a word global features because of their unique aspect.
2. the fundamental downward features forming the generic silhouette

We chose to use some global features for words to generate under-lexicons, these features permit us to define the general silhouette; notion that can be determined for each word of the lexicon with the following features:

– *diacritical signs*: points above the letters « ba » ; « ta » et « tha »

– *physical structure* : composed of the most stable downward areas (features);

– *physical length*: number of fundamental downward features.

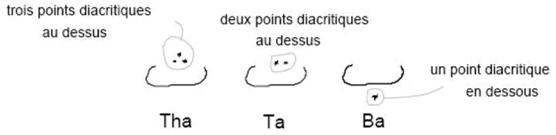

Fig. 10 . diacritical points of Arabic characters and their stable area

Using this complementary information aims at improving recognition, while correcting mistakes of recognition (confusion or groups of letters).

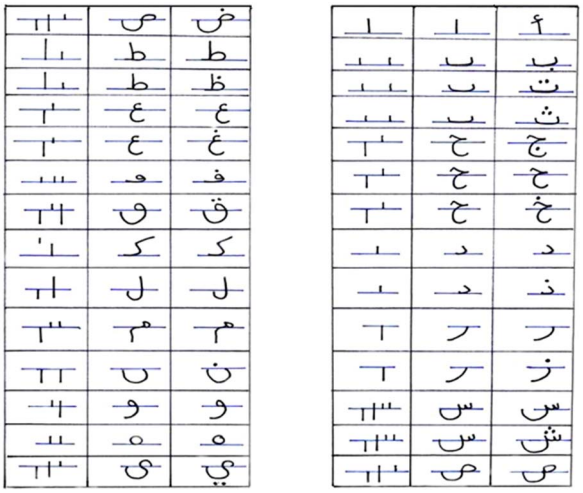

Fig. 11 . presentation of two levels of reduction

Using a representation by silhouette instead of a complete lexicon at the level of a recognition system, does not change anything at the level of the understanding but improves its performances. [8]

C. *Presentation of the approach of lexicon representation*

We present an approach of flexible lexicon modeling. The principle of this approach is: to use a simple representation of words based on global information. This representation can be defined from a handwritten word or directly on the basis of a character chain because we want it most independent possible of the writing style. During the phase of lexicon modeling, words of the lexicon are grouped according to their describers, forming thus under-lexicons indexed by a describer of under-lexicon.

During the phase of recognition, a describer of the word to recognize is extracted from its general silhouette, which permits to select under-lexicons having the nearest describers[12].

B. *Classification of words by generic silhouette*

The main idea of this representation is to group words having near physical structures and lengths in the same under-lexicons. The upward, downward and long features are regarded as prominent features.

In Arabic writing, the prominent features are more robust (independent of the writing style) and more easily detectable (because passing the body of the word significantly) than the median features. The physical structure of words is therefore compressed in order to be steadier and to form thus a generic silhouette: all prominent features are represented, but several successive median features are replaced by only one. According to the style of the writing (script, cursive or mixed) some words can be represented by several generic silhouettes and belong thus to several under-lexicons.

longue -ur	silhouette générique	longueur physique	structure physique	mot manuscrit
[4-11]		8		العربة
[4-11]		9		الورقة
[4-11]		11		المدرسة
[4-11]		9		العيادة

Fig. 12. Silhouette générique (les sous lexiques)

To compensate the loss of information produced by the method of compression, the generic length of the word (based on the number of downward features) is added to characterize the generic silhouette. Besides, the presence or absence of diacritical sign is taken into account. These three global features permit to organize the lexicon in under-lexicons. The visually near words (having the same generic silhouette, near generic lengths and absence or presence of diacritical signs) are regrouped in the same under-lexicons.

Le The describer of a word is represented then by a triplet (signdiacr, [lmin, lmax], S) and permits to group in an under-lexicon all words having the same generic silhouette S, of which the number of features is between lmin and lmax, and having the same value for signdiacr (0 for the absence of diacritical sign and 1 for the presence of at least one sign). A certain overlap of under-lexicons is introduced by the choice of generic length intervals to strengthen the stage of selection. Besides, some words are present in several under-lexicons or two characters having the same generic silhouette such as ص and ي.

V. EXPERIMENTATIONS AND ASSESSMENTS

Our objective is to integrate lexical knowledge within an embedded system of recognition, to model the lexical treatments to make them adaptable to the contexts of use and the evolutions

978-1-5090-6778-7/17/$31.00 ©2017 IEEE

of the system, while taking account of the constraints of memory capacity and computing time. For that, we proposed a specific modeling of the lexicon integrated for a reduction. We have the various experimental results allowing to evaluate the stages of the treatment. We detail the lexical resources which we have for these experiments.

The corpora of training and test are extracted from the magazine "Science and technology". We used 99948 words for the training and 4857 words for the test. The vocabulary is consisted of the 1126 words (including the unknown virtual word noted UNK) most frequent. Various methods of smoothing were used: Absolute, Kneser-Ney and Witten-Bell Discounting. In all the results presented below, the models are evaluated in term of perplexity (P).

Models of language learned on crudes corpora

The first developed model is calculated starting from N-grams without any pretreatment of the texts. Models of order 2,3 and 4 were thus learned. The table 1 present the values of perplexity obtained with UNKs.

According to this table, one notices well that:

- the best method of smoothing is that of Witten-Bell (lowest perplexity),
- formalization (representation) with the generic silhouettes always has a higher perplexity (see Fig 16).
- The model of order 2 always gives best perplexity.

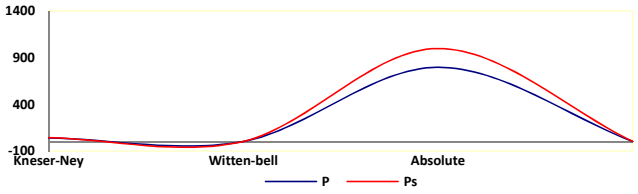

Fig. 13 . Perplexity of the model of order 3 (Modeling by generic silhouette).

A. *Models of language learned on corpora from morphemes (segmented text)*

An Arab word consists of a sequence of morphemes according to the diagram préfixes*-radical-suffixes* (* indicate zero or several occurrences of morphemes) [7]. To improve the performances of the models N-grams, the first pretreatment consists in segmenting partially the constituent words the corpora of training and test, and to rather calculate thus models containing morphemes than of words. By partial segmentation we understand to separate to it (S) prefix (S) from the remainder of the mot. Of the Arab examples of words and their segmentations are given in table 3. Table 4 lists the whole of the prefixes in which we were interested.

word	prefixes	Remain word
الولايات	ال	ولآيات
وليقوم	و, ل	يقوم

Fig. 14. Arab examples of words and their segmentation

The values of perplexity given in the tableau5 show that the models learned on such corpora are definitely more powerful than those calculated. These results prove thus that the segmentation of the text is an essential stage in the modeling of the Arab language[13].

B. *Models of language learned on bulkier corpora*

The corpora of training and test are extracted from the magazine "Science and technology". We used a corpus of training of 37 Mega of size and 10584 words for the test. The vocabulary is consisted of the 293601 words (including the unknown virtual word noted UNK) most frequent.

N	Kneser-Ney		Witten-Bell		Absolute		WithSmoothing	
	P	PS	P	PS	P	PS	P	PS
3	21.3873	21.3879	4.34133	4.34135	5.23891	5.23893	19.6911	19.6923

Fig. 15. Perplexity of the model of orders 3 calculated with UNKs on a larger corpus.

According to the following figure, one notices that the larger the corpus of training is, the more modeling by generic silhouettes approaches the representation by characters.

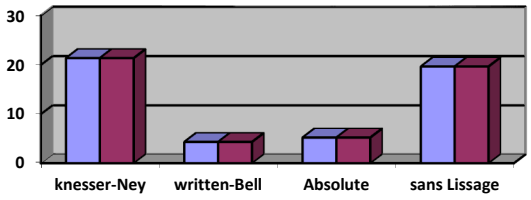

Fig. 16. Perplexity of the model of orders 3 calculated with UNKs on a larger corpus.

The experimental phase proved that modeling by generic silhouettes gives close results and in certain cases better than those led by a traditional modeling, which proves the originality and the relevance of our approach recommended.

REFERENCES

[1] A.BENSEFIA, T. PACKAGE, L HEUTTE, Identification and checking of the script writer in handwritten documents, *Treatment of the Signal*, vol. 22, No 3, pp.249-259, 2005.

[2] J.C. BEZDEK, *Pattern recognition with fuzzy objectifies function algorithms*, Plenum Near, 1981.

[3] J. - P. CRETTEZ, has set off handwriting families: style recognition. Acts of *ICDAR' 95*, vol. 1, pp. 489-494, 1995.

[4] F. GRIESS, S. CONNELL, On-line signature Checking, *Pattern Recognition,* vol. 35, n°12, pp. 2963-2,2002.

[5] KHOLMATOV, B. YANIKOGLU, Identity authentication using improved online signature checking method, *Pattern Letters Recognition*, vol. 26, pp. 2400-2408, 2005.

[6] A. HADDAD, (2004). Un système de génération automatique de dictionnaires linguistiques et thématiques de la langue arabe. Mastère en informatique, Ecole Nationale des Sciences de l'informatique, TUNISIE.

[7] H. MOUSSA, (1973). Statistical study of Arabic roots in moijam arous. Kouyet .

[8] CHANOD, J.-P et TAPANAINEN,P. (1995) : "Creating a tagset, lexicon and guesser for a french tagger " In Proceedings of EACL SIGDAT workshop on From Texts To Tags: Issues In Multilingual Language Analysis.

[9] Carbonnel. (2005) : Sabine Carbonnel ; Intégration et modélisation de connaissances linguistiques pour la reconnaissance d'écriture manuscrite en-ligne. Thesis.

[10] H. HABAILI, (1976). Contraintes de structure morphématique en Arabe, DEA en linguistique, Canada, université de Montréal.

[11] MERIALDO B, (1995):"Modèles probabilistes et étiquetage automatique" TAL, volume 36, 1995.

[12] R ZAAFRANI,(2004). Un dictionnaire électronique pour apprenant de l'arabe (langue seconde) basé sur corpus. JEP-TALN 2004, Fès, Maroc.

[13] T.SAIDANE, A.HADDAD, M.ZRIGUI, Pr. M. BEN AHMED, (2004). Réalisation d'un système hybride de synthèse de la parole arabe utilisant un dictionnaire de polyphones JEP-TALN 2004, Fès, Maroc.

[14] Haffar, Nafaa, Mohsen Maraoui, Shadi Aljawarneh, Mohammed Bouhorma, Abdallah Altahan Alnuaimi, and Bilal Hawashin. "Pedagogical Indexed Arabic Text in Cloud E-Learning System." International Journal of Cloud Applications and Computing (IJCAC) 7, no. 1 (2017): 32-46.

[15] Haffar, Nafaa, Mohsen Maraoui, and Shadi Aljawarneh. "Use of indexed Arabic text in e-learning system." In Engineering & MIS (ICEMIS), International Conference on, pp. 1-7. IEEE, 2016.

[16] O. Meddeb, M. Maraoui, and S. AlJawarneh: «Hybrid modeling of an OffLine Arabic Handwritting Recognition System AHRS», in International Conference on Engineering and MIS ICEMIS 2016, Agadir, Morocco, 22-24 September, 2016.

978-1-5090-6778-7/17/$31.00 ©2017 IEEE