A Global Online Handwriting Recognition Approach Based on Frequent Patterns

Chekib Gmati
LR-SITI Laboratory
National Engineering School of Tunis
El Manar University
Tunis, Tunisia
chekibgmt2007@gmail.com

Hamid Amiri
LR-SITI Laboratory
National Engineering School of Tunis
El Manar University
Tunis, Tunisia
hamidlamiri@gmail.com

***Abstract*—In this article, the handwriting signals are represented based on geometric and spatio-temporal characteristics to increase the feature vectors relevance of each object. The main goal was to extract features in the form of a numeric vector based on the extraction of frequent patterns. We used two types of frequent motifs (closed frequent patterns and maximal frequent patterns) that can represent handwritten characters pertinently. These common features patterns are generated from a raw data transformation method to achieve high relevance. A database of words consisting of two different letters was created. The proposed application gives promising results and highlights the advantages that frequent pattern extraction algorithms can achieve, as well as the central role played by the "minimum threshold" parameter in the overall description of the characters.**

Keywords-frequent features; mining frequent patterns; spatio-temporal relations; minimum threshold; online handwriting recognition

I. INTRODUCTION

The performance of the online handwriting recognition process is considered as an important characteristic in mobile devices [1]. Further, it could be a source of interesting information in several scientific fields. Indeed, in the field of medical data processing, a lot of tests have been done concerning the analysis of the muscular and neurological assessment of patients [2], such as the analysis of the relationship between learning motor skills, handwriting and reading [3]. Other research focuses on the detection of alcohol intoxication [4] or the analysis of the writing performance of dysgraphic children [5], biometrics, graphology, etc. Technological developments used in the conception of electrical devices like tablets, mobile phones and touch-sensitive notebooks demonstrate the need of human interactions using handwriting recognition algorithms [6, 7]. The online handwriting signal is a trajectory of a pen and could be represented by a sequence of coordinates of points ordered in time. This description of the signal highlights the spatio-temporal aspect and the intervention of the writer. Indeed, the variability of the writing by the same or several writers leads us to think extracting invariant and relevant characteristics that neglect parasitic and useless data. In this sense, a uniform description seems to be ideal to achieve this objective. However, in case of a signal carrying temporal information through sequential points, an elementary modeling is essential. Afterward, we pass from an elementary modeling to a global description through the extraction of frequent features.

II. RECOGNITION SYSTEM ARCHITECTURE

Online handwriting recognition is a four step process (Figure 1): online acquisition of the signal from an input device, pre-processing, primitives extraction and finally classification and recognition. The key step in this process is feature extraction. The nature of the handwriting is simply a temporal sampling applied to the incoming signal during the movement of the stylus tip on a touch screen; the result is a sequence of coordinates (x, y) of points.

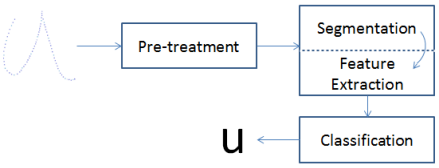


Fig. 1. The handwriting recognition process.

These points cannot be equidistant because the speed varies during writing. Their positions and the relationships that may exist between them depend on the movements of the writer. Between the sampling frequency and the writer's movement, several points will be recorded and can be considered as outliers in relation to all the points representing the object. Moreover, pre-processing (raw data normalization and smoothing) is essential in solving this problem. Actually, this step prepares the data in an adequate way and makes them appropriate for the rest of the process. Unnecessary and duplicated pixels as well as small arcs and hooks and all data that can represent noise, are eliminated. Then the character can be resized and centered to finalize the normalization phase. In addition, interpolation is used to recover missing pixels and detect sharp points to complete the relevant characteristics of the character. These techniques are further investigated in [8-10]. However, authors do not take into account the case where

some brackets and other delated data in the preprocessing step may belong to the specific characteristics of the character. Indeed, the writer can add to the object ligatures and brackets that will be considered as part of the model of the character and therefore they can contribute to the overall recognition process.

Before starting feature extraction, a segmentation step is applied to contribute directly in the modeling of the signal. In [11], authors presented a method to modify the dispersion of the points on the trajectory which tends to be concentrated on the areas where the speed of writing decreases. This is done based on spatial frequency analysis. Authors in [12] proposed a segmentation method that is however found unable to divide connected objects. Authors in [13] helped to address this difficulty by introducing the Freeman's coding, and by referring to the connection points as two opposing strokes. But this method is not applicable on complicated objects that we may encounter in online handwriting. On the other hand, several approaches adopt the segmentation-recognition method to solve the problem of handwriting online recognition [14]. Several approaches have been proposed in the literature based on online segments to extract satisfactory structural characteristics [15, 16]. Geometric feature extraction methods are applied such as line segment directions, line length [17], line order, spatial relationship between strokes, distances between consecutive points [18], tangents [19] and connection angles [20, 21]. Authors in [22] proposed an approach based on Delaunay triangles to describe the line segments. This makes it possible to select more "significant" groups (i.e. triangles) to represent the global characteristics and to use the temporal and topological characteristics of the handwritten forms. Other approaches are based on the Freeman code to code the direction of writing. An approach proposed in [21] is based on the Levenshtein distance computation. The characteristics are generated from a quantization of the contour. This quantification takes place in 8 zones with respect to the center of the character. It will then cover 360 degrees around the center. The contour will be coded according to this quantization method, and then two features will be added to the generated code chain. The first characteristic represents the value of the horizontal distance between the coordinate x of a point at time t and the coordinate of the starting point, respectively. The second characteristic represents the value of the vertical distance between the coordinate y from the same point at time t and the coordinate of the starting point. Relations between two points on the contour, distant in time, can be considered as characteristic, which would be added to the characteristic vector in relevance [22].

Feature extraction is a crucial step in the recognition process as it directly affects the efficiency of classification algorithms. Several classification methods have been proposed for online handwriting recognition such as neural networks, fuzzy logic, hidden Markov models (HMM), etc. [23]. HMM is a very popular classification method in online handwriting recognition systems. In [25], authors used an ergodic standard model with finite explicit statements (DHMM) for recognition of Arabic characters online. Each character is represented by a sequence of radial distances, and then a time distribution matrix in each state is defined. State duration modeling by DHMM improved discrimination by up to 12% compared to conventional HMMs. The disadvantages of HMM lie in the huge learning time and computing cost. A comparative study was carried out in [26] between HMM and SVM, a new feature extraction method based on the discrete wavelet transform. The results showed a very low learning time when using SVM compared to that of HMM. Authors of [27] used a Bayesian network algorithm for learning and adopted an approach based on positional relationships between strokes. This induces a spatial dependence between strokes, which represents a much better characteristic than the geometric one, but it is not adopted for online recognition. The modeling is done based on a hierarchy of points and strokes. The relationships between strokes and between points are explicitly and statistically modeled. In [30], authors introduced a method based on grammars from a hierarchical structure of handwritten kanji characters and used the HMM approach in the same context [31]. The method proposed in [28] is a hybrid approach that combines HMM and multilayer perceptron neural networks. The input signal is segmented in a continuous line, a characteristic vector is generated by the multilayer perceptron neural network based on ten characteristics. This generated vector is better adapted for the HMM, the approach was applied on the ADAB database [29].

III. FREQUENT ITEMSETS MINING

Data mining is a technique that extracts frequent patterns and it is a step of knowledge discovery in databases (KDD). The purpose of this step is to analyze the raw data, to extract useful information and to help make decisions [17]. In the knowledge discovery process, we found several steps that are proportional to the steps present in the process of online handwriting recognition systems. The first step, called data selection, is to clean the data to improve its qualities. This improvement also concerns the modification of the data to make them suitable for data analysis [15, 16]. Then, a cleaning step can be performed to eliminate redundancy, inconsistent or incorrect values. Regarding the preprocessing step, it can be applied to prepare the data for a transformation step, in order to refine the data quality. After this, we reach the stage of data mining, this step is crucial, because the quality of the generated models directly affects the efficiency of the system and influences the final decision.

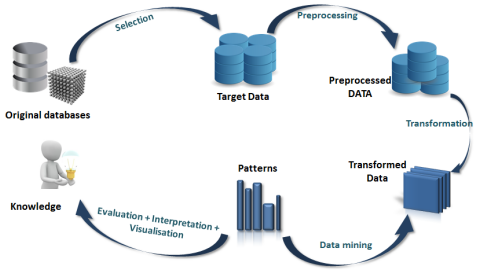


Fig. 2. KDD process.

A. Definition

Let I be a set of items $I=\{i_1, i_2, i_3, \ldots, i_m\}$ and D a set of transactions $D=\{t_1, t_2, t_3, \ldots, t_n\}$, where m is the number of

all items and n is the number of all transactions. Let X be another itemset of I, X⊆I and T a transaction of D, with T_X the set of transactions containing X [18]:

$$T_X = \{T \in D / X \subseteq T\} \quad (1)$$

Note that the support of X, *supp*(X), is only the ratio of the number of transactions containing X and the total number of all transactions in the database D.

B. Minimum Threshold

To be able to generate frequent itemsets, an essential parameter must be taken into account, it is a minimum threshold called minimum support and is named as minsupp. Indeed, an itemset is defined to be frequent if: $supp(X) \geq minsupp$.

C. Anti-Monotony

The main property that allows an optimal pattern extraction is the property of anti-monotony. Extraction techniques rely heavily on anti-monotonicity property to optimize extraction by reducing the number of items and candidates that are likely to be frequent [18]: Let $X^{'}$ and $X^{''}$ be two different itemsets, if $X^{'} \subseteq X^{''}$ then $supp(X^{'}) \geq supp(X^{''})$. Therefore, if $X^{'}$ is not frequent then $X^{''}$ is not frequent. This property was very useful when generating frequent itemsets. We will present two methods of extraction of the frequent pattens, the frequent and closed itemsets and maximal itemsets:

- Closed patterns: a pattern $X^{'}$ is defined as closed if no other pattern $X^{''}$ exist and has the same support of $X^{'}$, with $X^{'} \subseteq X^{''}$ [19].
- Maximal patterns: a pattern $X^{'}$ is maximal if no other pattern $X^{''}$ exists and is frequent, with $X^{'} \subseteq X^{''}$ [20].

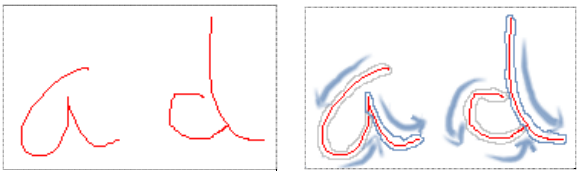
IV. Proposed Method

A. Data Selection and Preprocessing

We created a new dataset that contains the two characters. This dataset contains repetitive data that should be eliminated because it represents noise. Saving the same coordinates of a point for multiple times can produce a noise. This is due to the phase shift which can occur between the speed of the writing and the recording or sampling frequency of the signal. On the other hand, we will not remove small hooks because we want to prove we can achieve a good recognition rate while reducing the preprocessing step. Moreover, we want to prove that the characteristics relating to the sample still exist without eliminating strokes of the raw "form" of the character. Besides, to be able to describe the signal, we segment the trajectory into elementary strokes. Each of these elementary strokes is composed of 4 points in the chronological order of writing.

B. Extraction of Frequent Data

In the feature extraction step, the goal is to achieve a relevant representation of the object. In our approach we aim at extracting the frequent data, which are therefore dominant, in each object class of the database. Indeed, what we want to say is that we must extract the characteristics of the combination of all the letters of each word. This representation is interested in the relevant and discriminating part of all the data available in the class of each word, since there are several samples for each two-letter word. Figure 3(b) shows the strokes that share common characteristics. In fact the most frequent, or dominant features, will be extracted from the base of the samples of the same word, but also from the word itself.



(a) (b)

Fig. 3. (a) Two-letter word: 'ad', (b) Strokes that have common characteristics in a single word are surrounded by the same color.

To apply the extraction of frequent patterns we must group the calculated characteristics into three classes. Regarding major and minor axes, they will have the class "small", "medium" and "large". The directions will also have three classes: "up", "down" and "horizontal". Similarly, for the positions of the elementary strokes, they will have the classes "right", "left" and "continual". These classes are retrieved by computing the threshold values delimiting the groups after applying a clustering algorithm (kmeans). Thus, after the application of closed and maximal pattern extraction methods, we will obtain frequent itemsets of various sizes representing each object in the database. We must then normalize these vectors to obtain a matrix that we will compress to have a characteristic vector. We apply the standard deviation in a vertical manner to each matrix generated, i.e. according to the columns.

V. Results and Interpretations

The database of words we collected consists of several combinations of two Latin letters. The input device that we used was a graphics tablet. Recordings of the coordinates of touch points with the touch surface were performed every 10 milliseconds from the first pen-down. In Tables I-III, we present the characteristic and the classification algorithms that we used to test the relevance and efficiency of closed and maximal patterns in the description of objects bearing spatial and temporal information. We have varied the value of the minsupp to highlight the frequent itemsets from the point of view of relevance. In Figures 4 and 5, we get a peak when minsupp is equal to 0.55 or 55%, i.e. that we have obtained frequent motives of more than 55% and more relevant to the data set. On the other hand, the use of a closed pattern extraction algorithm (Charm) showed a better performance than that of the maximum pattern extraction algorithm (Charm MFI). Similarly for Figures 6 and 7, we note that the recognition rates exceed the previous results with a minimum of 50%. The values of the minimum thresholds are very close (50% and 55%), which demonstrates the validity of our approach regarding the relevance of frequent characteristics

that can be discriminatory. Moreover, closed patterns are more efficient than the maximal one; this is confirmed through the trend curves in the figures.

TABLE I. DESCRIPTION OF CHARACTERISTICS

Features	Description	usefulness
Major and Minor Axes	These are the values of the axes of an elliptical box applied to each elementary stroke.	These values give us the characteristics of the elementary strokes at the length and curvature level.
Directions	The direction relative to the vertical axis.	We consider it a major feature
Position of elementary strokes	This is the position of each elementary stroke compared to the preceding one.	Gives information on the orientation of writing while preserving invariance with respect to rotation and translation.

TABLE II. CORRECT CLASSIFICATION RATE (CHARM ALGORITHM)

Threshold Minsupp	Multilayer Percerptron	SVM	K-nearest neighbor	Naive Bayes
0.20	70.620	50.340	63.630	54.540
0.25	68.530	57.340	65.730	57.340
0.30	74.120	62.230	69.930	57.340
0.35	76.920	65.730	67.123	64.330
0.40	71.320	63.630	66.433	66.430
0.45	75.520	69.230	65.734	67.830
0.50	79.020	62.930	79.021	62.230
0.55	69.230	73.420	63.630	71.320
0.60	73.020	60.130	65.030	60.130
0.65	65.030	53.850	60.830	50.340

TABLE III. CORRECT CLASSIFICATION RATE (CHARM MFI ALGORITHM)

Threshold Minsupp	Multilayer Percerptron	SVM	K-nearest neighbor	Naive Bayes
0.20	58.741	51.748	61.538	55.944
0.25	62.237	54.545	59.440	57.342
0.30	66.433	51.049	61.538	57.342
0.35	66.433	49.650	65.734	51.049
0.40	76.223	60.139	58.042	60.139
0.45	72.727	60.139	63.636	62.937
0.50	68.531	60.839	66.433	64.335
0.55	70.629	65.734	62.237	69.230
0.60	67.132	57.342	69.930	60.839
0.65	62.237	51.049	62.937	52.447

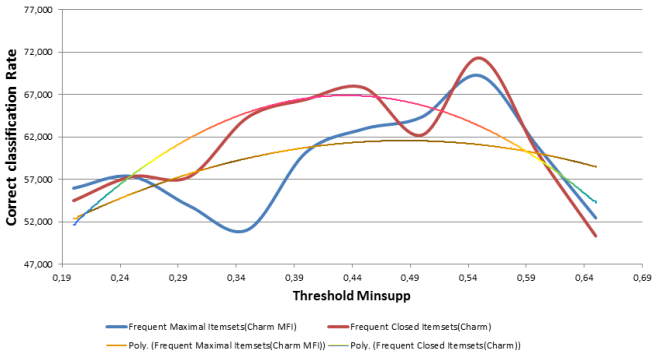


Fig. 4. Correct classidfication Rate (Naïve Bayes algorithm).

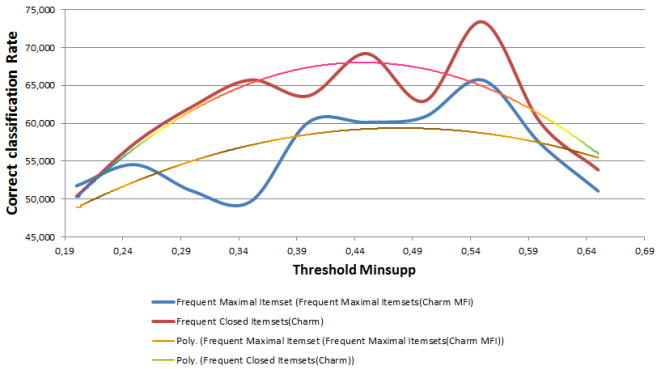


Fig. 5. Correct classidfication rate (SVM algorithm).

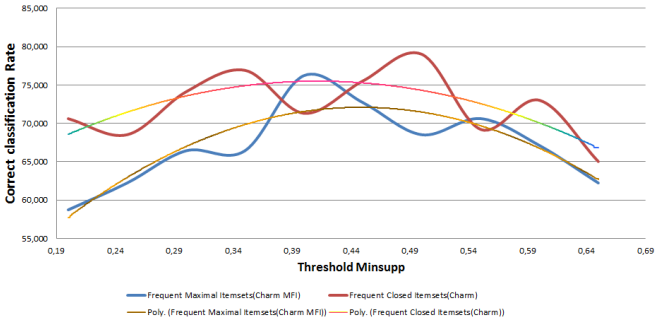


Fig. 6. Correct classidfication rate (Multilayer Perceptron Algorithm).

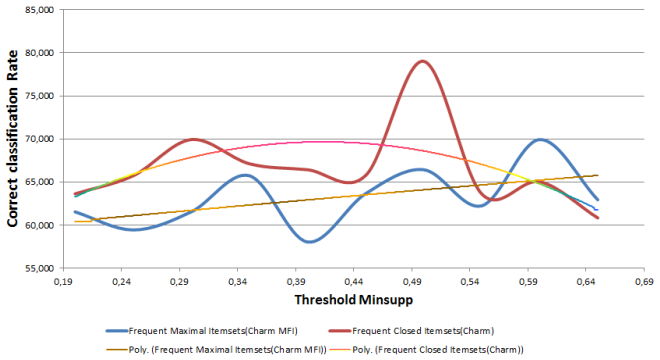


Fig. 7. Correct classidfication rate (K-nearest neighbor algorithm).

Increasing the value of the minsupp, data will be overlooked**.** When we get a peak at 50% or 55%, we can deduce that the noisy data are eliminated. The curves show that we can find relevant characteristics at values lower than 50%. It is true that this is not very remarkable, but it draws our attention to the cause of this phenomenon. The variations that we observe in all the figures show that there is an overlap between the discriminating characteristics and the characteristics that represent noise, and also show that the closed frequent motifs are more relevant because there is no loss in the generation of these patterns. We think the purpose of our proposed work is achieved, but these points need to be improved:

- To make the size of the elementary stroke adapted to the size of the object in a number of points to guarantee invariance in the case where the samples vary in size.
- Compression performed at frequent patterns by applying the standard deviation may not be sufficient to preserve the useful information that the frequent patterns generated.

VI. CONCLUSIONS

Through our approach, we generate characteristic vectors that contain dominant elements based on frequent pattern extraction algorithms. It has been proved that frequent closed patterns are more relevant with a minimum threshold value of 50% and 55%. The overall representation we have obtained can be remarkably improved if we focus on data fusion methods to preserve the useful information we have extracted. We noticed that the larger the database we have, the more the frequent patterns are discriminating. Thus, we are planning to prepare a massive database for handwritten character recognition online. Besides, we find that we need to introduce sequential pattern extraction algorithms to take advantage of temporal information that can be modeled through these algorithms.

REFERENCES

[1] I. Degtyarenko, O. Radyvonenko, K. Bokhan, V. Khomenko, "Text/shape classifier for mobile applications with handwriting input", International Journal on Document Analysis and Recognition, Vol. 19, No. 4, pp. 369-379, 2016

[2] N. Dounskaia, A. W. Van Gemmert, B. C. Leis, G. E. Stelmach, "Biased wrist and finger coordination in Parkinsonian patients during performance of graphical tasks", Neuropsychologia, Vol. 47, No. 12, pp. 2504-2514, 2009

[3] M. S. Julius, R. Meir, Z. Shechter-Nissim, E. Adi-Japha, "Children's ability to learn a motor skill is related to handwriting and reading proficiency", Learning and Individual Differences, Vol. 51, pp. 265-272, 2016

[4] J. Shin, T. Okuyama, "Detection of alcohol intoxication via online handwritten signature verification", Pattern Recognition Letters , Vol. 35, pp. 101–104, 2014

[5] V. Paz-Villagrán, J. Danna, J.-L. Velay, "Lifts and stops in proficient and dysgraphic handwriting", Human Movement Science, Vol. 33, pp. 381-394, 2014

[6] T. Deselaers, D. Keysers, J. Hosang, H. A. Rowley, "GyroPen: Gyroscopes for Pen-Input With Mobile Phones", IEEE Transactions on Human-Machine Systems, Vol. 45, No. 2, pp. 263-271, 2015

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989

[8] B. Q. Huang, Y. B. Zhang, M. T. Kechadi, "Preprocessing Techniques for Online Handwriting Recognition", Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007), Rio de Janeiro, Brazil, pp. 793-800, October 20-24, 2007

[9] M. A. Abuzaraida, A. M. Zeki, A. M. Zeki, "Online Recognition System for Handwritten Hindi Digits Based on Matching Alignment Algorithm", 3rd International Conference on Advanced Computer Science Applications and Technologies, Amman, Jordan, pp. 168-171, December 29-30, 2014

[10] M. E. Mustafa, H. A. A. Alshafy, "Characters' boundaries based segmentation for online Arabic handwriting", International Conference on Computing, Electrical and Electronic Engineering (ICCEEE), Khartoum, Sudan, pp. 306-310, August 26-28, 2013

[11] C. De Stefano, M. Garruto, A. Marcelli, "A multiresolution approach to on-line handwriting segmentation and feature extraction", IEEE 17th International Conference on Pattern Recognition (ICPR 2004), Vol. 2, pp. 614-617, 2004

[12] Y. Jiang, X. Wang, X. Ao, G. Dai, "Online Recognition of Handwritten Chemical Formula", 2nd Joint Conference on Harmonious Human Machine Environment. Hangzhou, China, pp. 111-115, 2006

[13] L. Zhao, H. Yan, G. Shi, J. Yang, "Segmentation of Connected Symbols in Online Handwritten Chemical Formulas", International Conference on System Science, Engineering Design and Manufacturing Informatization (ICSEM), Yichang, China, pp. 278-281, November 12-14, 2010

[14] M. Cheriet, N. Kharma, C. Liu, C. Suen, Character Recognition Systems: A Guide for Students and Practitioners, John Wiley & Sons, 2007

[15] O. Maimon, L. Rokach, Data Mining and Knowledge Discovery Handbook, Springer Science+Business Media, Inc, 2005

[16] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996

[17] S. Mitra, T. Acharya, Data Mining Multimedia, Soft Computing and Bioinformatics, John Wiley & Sons, 2003

[18] R. Agrawal, T. Imielinski, A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", in: SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pp. 207-216, ACM, 1993

[19] M. J. Zaki, C.-J. Hsiao, "ChARM: An efficient algorithm for closed itemset mining", in: 2002 SIAM International Conference on Data Mining, pp. 457-473, SIAM, 2002

[20] L. Szathmary, "Symbolic Data Mining Methods with the Coron Platform", PhD Thesis, Henri Poincaré University, 2002

[21] S. Dutta Chowdhury, U. Bhattacharya, S. K. Parui, "Online Handwriting Recognition Using Levenshtein Distance Metric", 12th International Conference on Document Analysis and Recognition, Washington DC, USA, August 25-28, 2013

[22] M. Mori, S. Uchida, H. Sakano, "Global feature for online character recognition", Pattern Recognition Letters, Vol. 35, pp. 142-148, 2013

[23] S. Dewangan, P. K. Gupta, U. K. Sahu, I. K. Verma, "Realtime Recognition of Handwritten Words using Hidden Markov Model", International Journal of Technological Synthesis and Analysis, Vol. 1, No. 1, pp. 7-9, 2012

[24] V. Vuori, M. Aksela, J. Laaksonen, E. Oja, "On-line recognition of handwritten characters", in: Biennial Report, Laboratory of Computer and Information Science, Neural Networks Research Centre, Helsinki University of Technology, 2003

[25] N. B. Amara, A. Belaïd, N. Ellouze, "Utilisation des modèles markoviens en reconnaissance de l'écriture arabe : état de l'art", Colloque International Francophone sur l'Ecrit et le Document - CIFEd'00, Lyon, France, July, 2000

[26] K. P. Primekumar, S. M. Idiculla, "On-line Malayalam Handwritten Character Recognition using HMM and SVM", International Conference on Signal Processing, Image Processing and Pattern Recognition (ICSIPR), Coimbatore, India, February 7-8, 2013

[27] S.-J. Cho, J. H. Kim, "A Bayesian Network Approach for On-line Handwriting Recognition", in: Digital Document Processing. Advances in Pattern Recognition, pp. 121-141, 2007

[28] N. Tagougui, H. Boubaker, M. Kherallah, A. M. Alimi, "A hybrid MLPNN/HMM recognition system for online Arabic Handwritten script", World Congress on Computer and Information Technology (WCCIT), Sousse, Tunisia, June 22-24, 2013

[29] H. El Abed, M. Kherallah, V. Margner, A. M. Alimi, "On-line Arabic handwriting recognition competition: ADAB database and participating systems", International Journal on Document Analysis and Recognition, Vol. 14, No. 1, pp. 15-23, 2011

[30] I. Ota, R. Yamamoto, S. Sako, S. Sagayama,"On-line Handwritten Kanji Recognition Based on Inter-stroke Grammar", IEEE 9th International Conference on Document Analysis and Recognition (ICDAR 2007), Vol. 2, pp. 1188-1192, 2007

[31] F. Alvaro, J.-A. Sanchez, J.-M. Benedí, "Recognition of on-line handwritten mathematical expressions using 2D stochastic context-free grammars and hidden Markov models", Pattern Recognition Letters, Vol. 35, pp. 58-67, 2014