Feature Selection using F-score Method for Offline Arabic Handwritten Fragment Identification

Soraya Azzoug, Youcef Chibani and Mohamed Anis Djoudjai
Laboratoire d'Ingénierie des Systèmes Intelligents et Communicant
Faculty of Electronics and Computer Science
University of Sciences and Technology Houari Boumédiene,
32, El Alia, Bab Ezzouar, 16111, Algiers, Algeria,
{sazzoug,ychibani,ma.djoudjai}@usthb.dz

Abstract. Among the various means of document authentication, handwriting fragments represent one of the challenging tasks for automatic handwritten writer identification. Usually, the writer is represented by a set of features extracted for each handwritten fragment. However, feature components might be irrelevant due to their redundancy, affecting the pertinence of the feature vector. Hence, this paper proposes a feature-selection strategy based on a hybrid F-score method, performed genuinely for multi-class classification. The F-score is used conjointly to a classifier based-distance. According to a statistical analysis performed on F-score distribution during the training step, the number of pertinent feature components is deduced from the highest F-score value. The experimental evaluation performed on the well-known IFN/ENIT handwritten fragment dataset shows an improvement of the identification rate of 94.20% while reducing the size of the feature vector of 54%.

Keywords: Handwritten fragments, writer identification, LPQ, feature-selection, hybrid method, F-score.

1 Introduction

The development of an Automatic Handwritten Analysis System (AHAS) in offline context can be useful for many civilian applications such as forensic, digital libraries, and text retrieval from historical documents [1-2]. Generally, offline AHAS are performed using either the entire document, paragraph, line, word, or the character [3]. Recently, a novel approach in offline AHAS consists of using some fragments of the whole document image and then performing the writer-identification process. Besides, the use of handwritten fragments provides additional information about the writing style which is the local information.

Hence, the literature has recently reported some attempts, exploring the use of handwritten fragments for writer identification, using particularly textural feature-descriptors such as Local Phase Quantization (LPQ) [4], Local Binary Patterns (LBP) [5], and Run Length (RL) descriptor [6]. However, feature vectors generated from these descriptors can include irrelevant feature components due to their redundancies, which affects negatively the identification accuracy.

To handle this issue, various feature selection algorithms can be performed for selecting only pertinent features. Basically, three main approaches for feature-selection can be found: Filter, wrapper and hybrid. Filter models rely on applying various fast algorithms for ranking features according to an evaluation of their relevance such as Chi-square (CHI), Information Gain (IG) and T-Test based Feature Selection (TTFS) [7-8]. In contrast, wrapper models are based on generating feature sets *via* classifiers such as Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN) [9]. Although encouraging performances are achieved, both approaches cost usually significant memory and time execution due to their complexity. For better efficiency, hybrid algorithms have been introduced combining the two feature-selection approaches (i.e. filter and wrapper) [9]. It consists of employing, first, a filter evaluation function to preselect the most discriminating feature subspaces. Then, a classification step is performed on the previous selected subspaces for selecting the final subspace.

Hence, this paper proposes the use of the hybrid selection method, namely F-sore, for offline Arabic handwritten fragments identification. The evaluation of the proposed system on the ENIT-IFN dataset shows the improvement of performances after applying the proposed feature selection scheme. To the best knowledge, the use of F-score method has never been explored for Arabic handwritten fragment identification.

The remaining of this paper is structured as follows: Section 2 gives a review of F-score method. Section 3 explains the proposed methodology. In section 4, the obtained results are presented and discussed. Finally, the last section presents the conclusion and future works.

2 Review of F-Score Algorithm for Bi-class

Fisher-score or commonly called F-score, is one of the most widely used supervised feature selection methods. It is a simple but effective method to measure the discrimination between two sets of real numbers. Early works applying F-score were introduced by Yi-Wei Chen et al., [10] for face recognition problem. In writer identification context, F-score can be used for measuring the degree of feature's relevance, discriminating between two different writers.

Originally, the F-score has been designed for discriminating between two classes. Let a data sample represented by its feature vector $X_k = \{x_{k,1}, x_{k,2}, \dots, x_{k,N}\}$ of size N, $k = \{1,2, \dots., R\}$, where R is the number of samples. For Bi-class classification, a given feature vector of a sample can belong either to the positive class denoted $X_k^{(+)}$or the negative class denoted $X_k^{(-)}$. Hence, the F-score measure computed for a given feature component $x_{i,k}$ in Bi-Class classification namely F_{BC}, is defined as follows:

$$F_{BC}(x_{i,k}) \equiv \frac{\left(\mu_i^{(+)}-\mu_i\right)^2+\left(\mu_i^{(-)}-\mu_i\right)^2}{\frac{1}{R-1}\sum_{k=1}^{R^+}\left(x_{k,i}^{(+)}-\mu_i^{(+)}\right)^2+\frac{1}{R-1}\sum_{k=1}^{R^-}\left(x_{k,i}^{(-)}-\mu_i^{(-)}\right)^2} \quad (1)$$

$$\mu_i^{+} = mean_k^R\left(x_{k,i}^{+}\right) \quad (2)$$

$$\mu_i = mean_j^{W_{BC}}\left(mean_k^R\left(x_{k,i}^{j}\right)\right) \quad (3)$$

where W_{BC} represents the number of classes in the Bi-class context (i.e.$W_{BC} = 2$). $\mu_i^{(+)}$ and $\mu_i^{(-)}$ are the *mean* value calculated for the feature component $x_{k,i}$ of the rank i, belonging to the positive and the negative classes, respectively. While μ_i is the *mean* value calculated for the feature component of the rank i considering all the samples belonging to the set R. The numerator indicates the *inter-class* discrimination, while the denominator indicates the *intra-class* discrimination. Hence, the relevance of the feature is determined *via* computing the F_{BC}, i.e. the higher F-score is, the more the feature is relevant. Nevertheless, the Bi-class F-score (F_{BC}) formulation is not adapted to the targeted application, which is the writer-identification. Thus, a general formulation of F-score should be considered in the context of the multi-class classification. In the section, the extension of the F-Score for multi-class classification is proposed for writer identification using handwritten text fragments.

3 Proposed System

The proposed Offline Writer Identification System (OWIS) is composed of three main modules as depicted in Fig. 1: Feature generation, feature selection and classification. The details of each module are provided in the next sections.

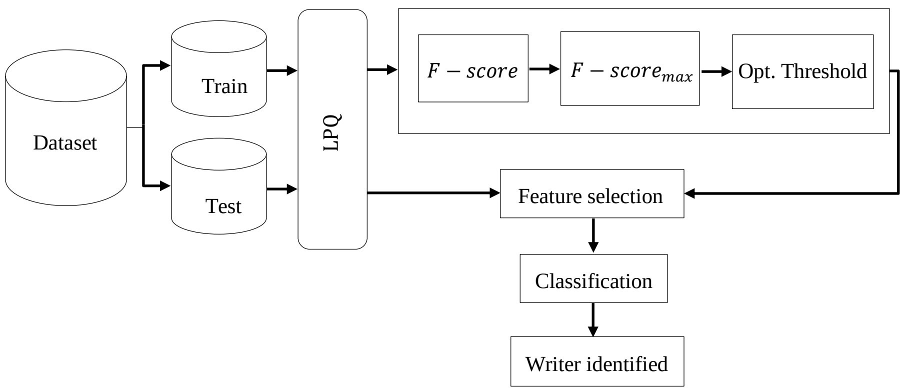

Fig. 1. Scheme of the proposed OWIS.

3.1 Feature generation

Among the various existing feature generation methods, the Local Phase Quantization (LPQ) is employed in this work for its efficiency in characterizing the text fragment, and for the fair comparison with the state-of-art. LPQ is based on quantized phase information of the 2D Discrete Fourier Transform (DFT). More precisely, the local phase information is extracted using the Short-Term Fourier Transform (STFT), computed at each pixel position x of the image $I(x)$, over a rectangular $M \times M$ neighborhood namely N_x, such as:

$$STFT(u, x) = \sum_{y \in N_x} I(x - y) e^{-j2\pi u^T y} = w_u^T \, I_x \tag{4}$$

where w_u is the basis vector of the 2D STFT at the frequency u. Besides, LPQ uses only four complex coefficients that corresponds to the following 2-D frequencies: $u_1 = [a, 0]^T, u_2 = [0, a]^T, u_3 = [a, a]^T, u_4 = [a, -a]^T$, where a is a scalar sufficiently small to satisfy $H(\mu_i) > 0$.

Therefore, the feature vector obtained via LPQ method namely $STFT_X$ can be represented as:

$$STFT_X = [\, STFT(\, u_1, x), STFT(\, u_2, x), STFT(\, u_3, x), STFT(\, u_4, x) \,] \tag{5}$$

The size of the obtained feature vector is $N = 255$. Further details about the mathematical formulation of LPQ can be found in [11].

3.2 Feature selection

The proposed feature selection is performed through two steps: Computing the F-score on multi-class classification and selecting optimal features using a threshold.

Computing the F-score for multi-class classification. Let S^j be a set of R handwritten fragments $(s_1^j, s_2^j, \dots, s_R^j)$ of the class j, such that $j = 1,2, \dots, W_{MC}$, where W_{MC} represents the number of writers. Let the feature vector $X_k^j = \{x_{k,1}^j, x_{k,2}^j, \dots, x_{k,N}^j\}$ of size N, generated from the handwritten fragment s_k^j, such that $k = 1,2, \dots, R$, belonging to the class j. Hence, the F-score measure computed for a given feature component $x_{k,i}^j$ where $i = 1,2, \dots, N$, in a Multi-Class classification namely F_{MC}, is defined as follows:

$$F_{MC}\left(x_{k,i}^j\right) = \frac{\sum_{j=1}^{W_{MC}} \left(\mu_i^j - \mu_i\right)^2}{\sum_{j=1}^{W_{MC}} \left(\frac{1}{R-1} \sum_{k=1}^{R} \left(x_{k,i}^j - \mu_i^j\right)^2\right)} \tag{6}$$

$$\mu_i^j = mean_k^R\left(x_{k,i}^j\right) \tag{7}$$

$$\mu_i = mean_j^{W_{MC}}\left(mean_k^R\left(x_{k,i}^j\right)\right) \tag{8}$$

where μ_i^j represents the *mean* value computed for the feature component x_{ki}^j of the rank i, belonging to the writer j. While μ_i is the *mean* value calculated for the feature component of the rank i considering all the fragments R and all the writers W_{MC}.

Selecting the threshold. After calculating the F-score of each feature component ($x_{k,i}^{j}$) of the feature vector (X_k^j), the next step consists of defining a convenient threshold value namely τ with which irrelevant feature components will be deleted.

Hence, the first step consists of taking the $median$ F-score value namely δ_i, considering all F-score values (i.e. $W_{MC} \times R$) computed for each feature component ($x_{k,i}^{j}$), belonging to the same rank i, taken from the feature vector of the fragment k, and associated to the writer j, such that:

$$\delta_i = median_j^{W_{MC}}\left(median_k^R\left(F_{k,i}^j\right)\right) \tag{9}$$

Consequently, a set $\Delta = \{\delta_1, \delta_2, \dots, \delta_N\}$ is obtained, where N the size of the feature vector is (X_k^j). For selecting the optimal threshold value τ, the $maximum$ metric is used as follows:

$$\tau = max_i^N(\delta_i) \tag{10}$$

Finally, the feature-selection rule denoted Γ, used for deleting irrelevant feature components, is defined according to the following equation:

$$\Gamma = \begin{cases} x_{k,i}^j \ \ accepted \quad if \quad i < N_\tau \\ x_{k,i}^j \ \ rejected \quad\quad Otherwise \end{cases} \tag{11}$$

where N_τ represents the corresponding number of necessary feature components for having the *maximum* F-score value τ. During the evaluation step, the same percentage corresponding to N_τ found during the training step, is kept and used straightforward for reducing the irrelevant feature components.

3.3 Classification

For comparing a query handwritten fragment represented by its feature *vector* $X_q = \{x_{q,1}, x_{q,2}, \dots, x_{q,N}\}$ of size N, against a reference handwritten fragment k represented by its feature vector $X_k^j = \{x_{k,1}^j, x_{k,2}^j, \dots, x_{k,N}^j\}$, belonging the writer j, the Hamming distance based-classifier [7] is applied for its efficiency, such as:

$$D_H(X_q, X_k^j) = \sum_i^N \left|x_{q,i} - x_{k,i}^j\right| \tag{12}$$

Finally, based on the D_H value, the query handwritten fragment represented by its feature vector X_q is assigned to the class label $y(X_q)$ through the following equation:

$$y(X_q) = argmax_j^{W_{MC}}(D_H(X_q, X_k^j)) \tag{13}$$

4 Experimental Results

4.1 Dataset description

For evaluating the proposed OWIS, the large Arabic handwritten dataset namely IFN/ENIT dataset [7] is used. It is the most widely used Arabic handwritten dataset proposed by Pechwitz et al., [11]. It comprises 2200 forms with more than 26 000 handwritten Arabic Tunisian town/village names collected from 411 different writers. All forms are scanned at a resolution of 300 dpi and are available as binary images.

4.2 Experimental protocol

In this work, the IFN/ENIT dataset is subdivided into two subsets: Subset 1 contains 2/3 of the entire IFN/ENIT dataset, which is used for training the OWIS. While, the Subset 2 contains the remaining 1/3 of the entire IFN/ENIT dataset, which is used for testing the proposed OWIS.

For a fair comparison, text fragments generated by Hannad et al., [7] are straightforward used in this work. The procedure is briefly reviewed, which can be performed through the following steps:

- Binarization of handwritten images *via* a global thresholding.
- Extraction of connected components from the handwritten image.
- Generation of small handwritten fragments by dividing each connected into a window size 100×100.

The performance of the proposed OWIS is evaluated using the Identification Rate (IR) expressed in percentage (%) as follows:

$$IR = 100 \times \frac{Number\ of\ correctly\ identified\ handwritten\ fragments}{Total\ number\ of\ handwritten\ fragments} \tag{14}$$

4.3 Experimental evaluation

For highlighting the effect of applying F-score on feature component values, Fig. 1 depicts the distribution of the normalized F-score values versus the number of feature components, following the same technique detailed in section 3 (Selecting the threshold). Hence, the obtained curve demonstrates clearly the usefulness of applying F-score for capturing the relevance of feature components. Indeed, it shows that F-score takes initially an oscillating values, until reaching the highest F-score value for the feature component number 139 ($N_\tau = 139$), which corresponds to the optimal number (i.e. $\sim 54\%$ of the initial feature vector size) of pertinent feature components considered for the classification step.

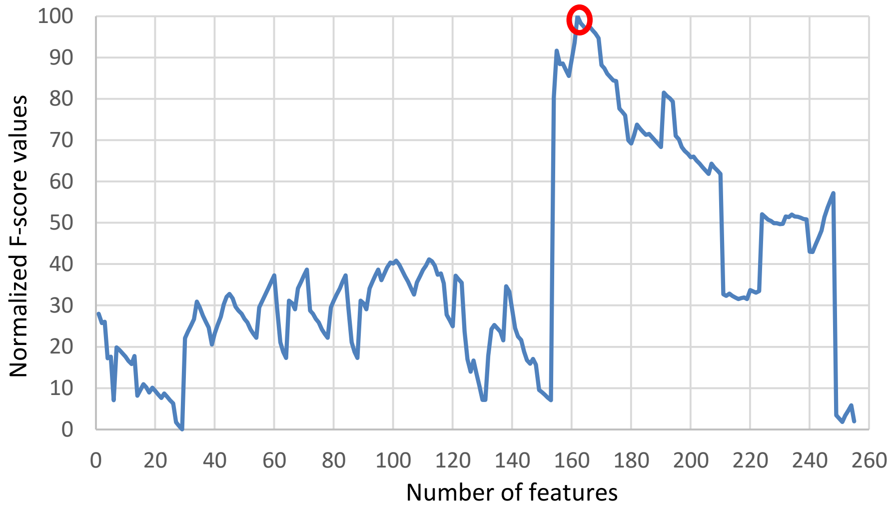

Fig. 3. Evolution of F-score versus the number of feature components.

The achieved results performed on the ENIT/IFN dataset, before and after feature-selection denoted FS, are depicted in Table 1. As can be seen, an improvement of the identification rate is reported when a reduction of 54% of the initial feature vector size is performed. Hence, the obtained results demonstrate the effectiveness of the proposed feature-selection scheme, based on the hybrid F-score method, for writer identification. In other words, the presence of irrelevant feature components contained into the feature vector influences negatively the quality of the handwritten fragment representation, and thus, the process of recognition (writer-identification).

Table 1. IR (%) obtained before and after feature selection performed on the ENIT/IFN dataset

Method	*N*	IR (%)
Before FS	255	93.91
After FS	139	**94.20**

4.4 Comparative Analysis

For a better evaluation of the proposed feature-selection method, a comparison against the-state-of-the-art works is performed using the same text fragments belonging to the dataset (ENIT/IFN). Table II summarizes the performance achieved by the last similar works reported in the literature using the LPQ descriptor.

Table 2. Comparison of performances achieved by the proposed work and the-state-of-the-art works on the ENIT/IFN dataset.

Works	#Features	Distance	FS	IR (%)
Abdi et al., [9]	240	Euclidean	No	90.02
Hannad et al., [7]	255	Hamming	No	94.89
Hadjadji et al., [6]	255	Hamming	No	94.40
Proposed work	139	Hamming	Yes	94.20

The best performance is achieved by Hannad et al., [7] namely 94.89% of IR. However, even the same feature descriptor is used (LPQ) for generating features, they used the entire feature components (255). In contrast, the proposed work achieves a close identification accuracy of 94.20% using only 139 feature components. Hence, the proposed writer identification scheme achieves a close performance with less complexity in time execution and memory space.

4 Conclusion

The objective of this paper aimed to explore the use of the F-score for feature selection in the context of the multi-class framework performed on the Arabic handwritten fragment identification. Based on the statistical analysis performed during the training step, a new thresholding scheme based on the F-score is proposed for selecting only the relevant feature components. The obtained results of the proposed Offline Writer Identification System (OWIS) demonstrate the effectiveness of using the F-score method for improving performances, even when a reduced number of feature components is used. F-score algorithm can be considered as a good alternative for feature-selection task, applied to the handwritten fragments of the ENIT IFN database.

As future work, an interesting work consists of applying the feature selection on a larger and a consistent feature vector, along with a robust classifier, for improving the identification accuracy. Furthermore, the use of F-score for selecting pertinent handwritten fragments (in addition to the feature-selection) can also be an interesting work for improving the effectiveness of the proposed OWIS.

Acknowledgement

This work was supported by the Direction Générale de la Recherche Scientifique et du Développement Technologique (DGRSDT) grant, attached to the Ministère de l'Enseignement Supérieur et de la Recherche Scientifique, Algeria.

References

1. Fajri Pratama S, Huoy choo Y, Muda A (2010) Feature Selection Methods for Writer Identification:Acomparative Study. International Conference on Computer and Computational Intelligence, Nanning, China.
2. Schomaker L, Bulacu M (2007) Text-independent writer identification and verification using textural and allographic features. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(4):701-717.
3. Bensefia A, Paquet T, Heutte L (2005) A Writer identification and verification system. Pattern Recognition Letters 26(13): 937-941.
4. Ojansivu V, Heikkila J (2008) Blur intensive texture classification using local phase quantization. International Conference in Image and Signal Processing ICISP 2008, pp 236-243.
5. (Ojala, Pietikäinen, & Mäenpää, 2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern Analysis and Machine Intelligence, 24(7): 971-987
6. Hadjadji B, Chibani Y (2018) Two combination stages of clustered one-class classifiers for writer identification from text fragments. Pattern Recognition 82: 147-162. doi:10.1016/j.patcog.2018.05.001.
7. Hannad Y, Siddiqi I, El Kettani M (2016) Writer identification using texture descriptors of handwritten fragments. Expert Systems With Applications 47:14-22.
8. Bulaco M, Schomaker L, Vuurpijl L (2003) Writer identification using edge-based directional features. The 7th International Conference on Document Analysis and Recognition, pp 937-941.
9. Abdi M N, Khemakhem M (2012) Arabic Writer Identification and Verification Using Template Matching Analysis of Texture. IEEE 12th International Conference on Computer and Information Technology ICCIT'2012, Chengdu, China, pp. 592 – 597.
10. Chen YW, Lin CJ (2006) Combining SVMs with Various Feature Selection Strategies. Studies in Fuzziness and Soft Computing, vol 207. Springer, Berlin, Heidelberg 207:315–324.
11. Pechwitz M, Maddouri S S, Margner V, Ellouz N, Amiri H (2002) IFN/ENIT-database of handwritten Arabic words. Proceedings of CIFED: Vol.2 pp.127-136.