 ISSN: 2579-7298

Bio-inspired Expert System based on Genetic Algorithm for Printer Identification in Forensic Science

Saad Mohamed Darwish[1], Hany M ELgohary[2]

a Department of Information Technology Alexandria University, Egypt.
[1] saad.darwish@gmail.com; [2] hany_elgohary44@yahoo.com*
** corresponding author*

ARTICLE INFO

Article history:
Received: 2018-05-11
Revised 2018-06-05
Accepted 2018-08-15

Keywords:
Printer identification,
Texture Analysis,
Genetic Algorithm,
Feature Selection,
KNN

ABSTRACT

Printer identification models are provided for the goal of distinguishing the printer that produced a suspicious imprinted document. Source identification of a published document can easily be a significant procedure intended for the forensic science. The arising problem is that the extraction of many features of the printed document for printer identification sometimes increases time and reduces the classification accuracy since a lot of the document features may come to be repetitive and non-beneficial. Distinct combinatorial collection of features will need to be acquired in order to preserve the most effective fusion to accomplish the maximum accuracy. This paper presents an intelligent machine learning algorithm for printer identification that adopts both of texture features formulated from gray level co-occurrence matrix of the printed letter "WOO" and genetic heuristic search to select the optimal reduced feature set. This integration aims to achieve high classification accuracy based on small group of discriminative features. For classification, the system utilizes k-nearest neighbors (KNN) to recognize the source model of the printer for its simplicity. Experimental results validate that the suggested system has high taxonomy accuracy and requires less computation time.

I. Introduction

Secure printings involve the strategy that printer outcome, named a document, is a successful way to distinguish several features of the printer. These kinds of features, which in turn are printer specific, can easily be utilized for document security. For example, in case of planned forgery, all of us ideally should certainly be ready to recognize the category of printer that was appointed to produce the document [1]. The importance of document security has increased due to the ease of counterfeiting or forgery of documents such as banknotes, official documents, bank check, visa, driver's licenses, and passports according to the development of computer and printing technologies. Therefore, the ability to embed and extract information in/form printed documents would be desirable for many security applications [2]. Published documents include features of the printing device based on the particular technique employed by producers for inserting the tagging element on the document [3]. The printer identification is closely related to various pattern identification and recognition techniques [4].

Researches into printer identification have been focused on two techniques, namely passive and active based on examining the printed document [5]. The passive technique involves characterizing the printer by finding intrinsic features in the printed document that are characteristic of that distinctive particular printer, model, or manufacturer's products. The intrinsic signature requires an understanding and modeling of the device mechanism and the development of analysis tools for the detection of the signature in the printed document with arbitrary content [6-9]. The passive technique is able to successfully determine printer types (e.g. laser, inkjet) and printer's makes and models based on some features such as banding frequencies, pattern noise features, geometric distortion, printer profiles, and texture features [10]. In general, texture features can be more easily

measured over small areas such as inside a text character and has also been shown to work across varying font type, font size, and printer consumable age with has high discrimination accuracy.

In contrast, the active technique embeds an extrinsic signature in the printed page. This signature is generated by modulating the process parameters in the printer mechanism to encode identifying information such as date of printing, the printer's serial number, and time of printing[5][11][12]. Active technique embeds traceable information into the document, being it imperfections in text or images, or microscopic tracking dots that encode the printer's serial number. These tracking dots are yellow in color printed on a white background, making them appear invisible to the naked eye. However, this technique is reportedly only used with the color laser printer, which limits its application dramatically. A large number of documents don't require color and may be printed using a grayscale facility. Most of printer identification systems in the forensic science employ the passive technique (intrinsic signature) because active techniques modify the printing process parameters that can induce unexpected printing quality. Intrinsic signature is tied directly to the electromechanical properties of the printer; so it is hard to forge or remove. [8][12]. Also, yellow dots contain encoded information the naked eye can hardly see it [13].

There are many challenges associated with the printer identification task for Arabic manuscript such as [1][11] [12][14]: (1) Arabic characters can have more than one shape according to their position in a word : isolated, begin of word, middle of word, and end of word. (2) Several variables can affect the performance: the type of paper, font type, font size, printer consumable age, work with multiple font sizes, and also different characters increases the complexity. (3) Because of the advancing technologies in the world, various image processing tools are available to forge the documentation easily and efficiently; so that the authentication of printed data is a big challenge. (4) Many features of the printed document for printer identification sometimes increase time and reduce the classification accuracy of the recognition system since some of the features may be redundant and nom-informative. An efficient method to solve these difficulties is by utilizing the genetic algorithm (GA) for feature selection. Feature selection can be defined as a process that chooses a minimum subset of features from the original set of features; so that the feature space is optimally reduced according to certain evaluation criteria [14]. GA is now widely applied in science and engineering as adaptive algorithms for optimizing practical problems based on principles of natural selection. Based on the concept of the best fitness value of a GA, optimal features can be easily achieved [15].

This paper focuses on the research of printer identification for Arabic alphabet, which is still a challenging research topic and not extensively explored by researchers. The work presented in this study tries to extract Gray Level Co-occurrence Matrix features (GLCM) from the printed letter " WOO " as it is one of the most used alphabets in the Arabic language, and is written completely in any position of the word. The system also explores the optimum feature subset by using bio-inspired feature selection technique. For classification, it employs the KNN classifier as one of the most famous neighborhood classifiers in pattern recognition. In this case, an easy and effective way to calculate the classification error rate is by the "leave one out "procedure. The classification accuracy of KNN is considered as the fitness function for GA.

The outline of the remainder of this paper is as follows. Section 2 describes some related work in the printer identification. Section 3 describes the proposed system. Section 4 summaries experimental results, and Section 5 concludes the paper.

II. Method

In this section, we will explain the framework of printer identification system based on image texture analysis. The system utilizes GLCM method for getting the features of a particular printer, then GA is adapted to select the optimal feature set to be used for classification. Fig. 1 shows the block diagram of our printer identification scheme. Each step will explain in details. We have collected our data from 10 different printers of different brands with a different model and serial number. After data collection, the documents are scanned at 1200 dpi with 8 bits/pixel (grayscale) because the high-resolution image appears crisper, and its texture will often be more clear and vibrant. Then all the features have been extracted from the isolated character "و".

Character Extraction

This step extracts the "و "character in the document image because it is one of the most used alphabets in the Arabic language. It can be noticed from Fig. 2 that different printers print this particular alphabet differently. This character was also chosen based on prior experiments to initially test the accuracy of the identification using a different character.

Fig. 1. Character "WOO" printed from different printers.

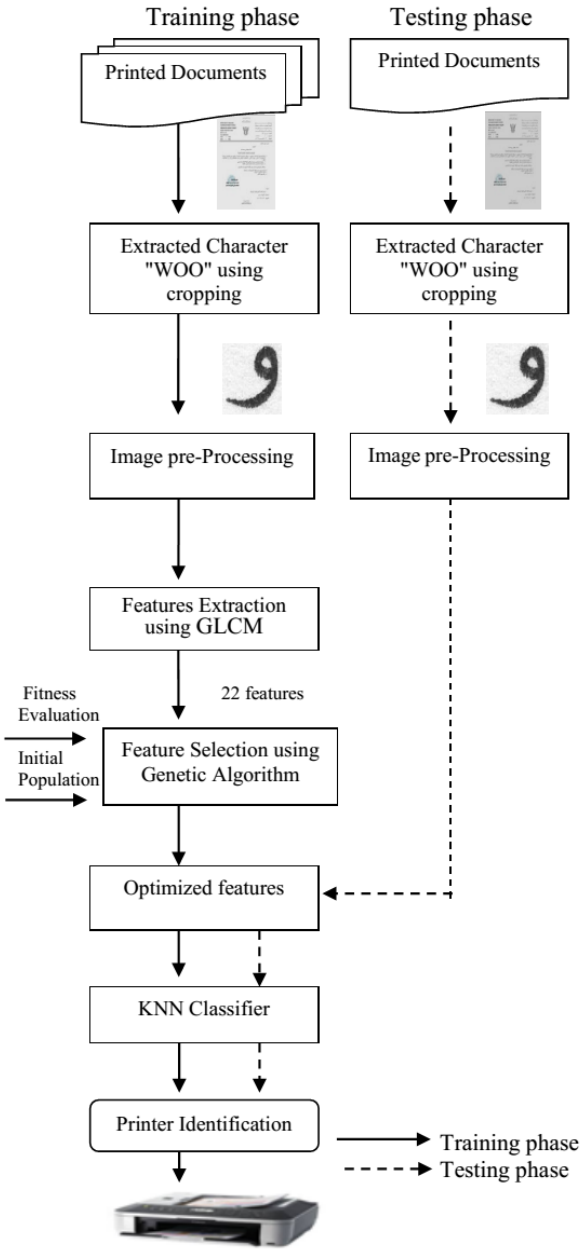

Fig. 2. System diagram of source printer identification scheme.

Image Pre-processing stage

The preprocessing stage is implemented both in the training and testing phases. The purpose of the pre-processing is to get the coordinates of the center of every character in an ideal form and prepares the document image for simple and easy features extraction step [5]. The preprocessing stage follows six steps [18][19].

a) ***Conversion to grayscale****: Since the attention is only in the grayscale image and not in its color, color information is irrelevant.*

b) ***Binarization****: The grayscale character is treated by a histogram-based binarization to produce a binary image that contains only 0's and 1's.*

c) ***Noise reduction****: Once the original image is binarized, the next step is to remove the noise from character image caused during scanning via median filtering method.*

d) ***Image cropping****: The binary image is segmented from the background to remove the white space surrounding the character using the segmentation method of vertical and horizontal projections.*

e) ***Rotation and width normalization****: The cropped image is scaled using bi-cubic interpolation to a constant width, keeping the aspect ratio fixed. The positional information of the character is normalized by calculating an angle θ about the centroid (x, y) such that rotating the character by θ brings it back to a uniform baseline. The character's size normalization is important because it establishes a common ground for image comparison. Herein, Taylor's maximization is used for normalization*

f) ***Thinning****: The goal of thinning is to produce a simplified, but the topologically equivalent image to assist in features extraction and classification.*

Feature extraction based on GLCM

We want to be able to determine a set of features that can be used to describe the output documents of the printer. The proposed system treats the scanned document as an "image" and uses image analysis tools to determine the features that characterize the printer. Each printer has different sets of banding features that are dependent upon brand and model. Banding features of a printer caused by electromechanical fluctuations and imperfections are relatively easy to estimate from documents with large mid-tone regions. However, it is difficult to estimate the banding features from the text. GLCM is a widely used texture analysis method especially for stochastic textures to find a feature or set of features that can be measured over smaller regions of the document such as individual text characters. GLCM has also been shown to work across varying font type, font size, printer consumable and printer age [8][12][21].

The GLCM is a tabulation of how often different combinations of pixel brightness values (gray levels) occur in an image. The advantage of the co-occurrence matrix calculations is that the co-occurring pairs of pixels can be spatially related in various orientations with reference to distance and angular spatial relationships, as on considering the relationship between two pixels at a time [21]. To generate a GLCM, first, we define the number of pixels in the ROI (region of interest), which is the set of all pixels within the printed area of the character. There are a total of 22 features that could be computed from GLCM that is calculattion as:

$$G(i,j) = \frac{p(i,j)}{\sum_{i=0}^{n}\sum_{j=0}^{n} p(i,j)} \quad (1)$$

where G is the normalized GLCM , n is the number of the GLCM elements, and $p(i,j)$ represents the number of occurrences of grey levels i and j within the window.

$$Contrast = \sum_{i=0}^{n-1} n^2 \sum_{j=0}^{n-1} G(i,j)(i-j)^2 \quad (2)$$

$$Dissimilarity = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} G(i,j)\left|i-j\right| \quad (3)$$

$$Homogeneity = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \frac{G(i,j)}{1+(i-j)^2} \quad (4)$$

$$Moment_{Angular} = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} G(i,j)^2 \tag{5}$$

$$Similarity = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \frac{G(i,j)}{1+|i-j|} \tag{6}$$

$$Entropy = -\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} G(i,j).\,Ln\,[G(i,j)] \tag{7}$$

$$Mean = \mu = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} iG(i,j) \tag{8}$$

$$Correlation = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \frac{G(i,j)\,[(i-\mu_x)(j-\mu_y)]}{\sigma_x \sigma_y} \tag{9}$$

$$\sigma_x = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} G(i,j)(i-\mu_x)^2 \tag{10}$$

$$\text{Probability}_{\text{Max}} = \max_i \max_j G(i,j) \tag{11}$$

$$Moment_{Diagonal} = \sum_{i=0}^{n-1} j \sum_{j=0}^{n-1} \left(0.5G(i,j)|i,j|\right)^{0.5} \tag{12}$$

$$Moment_{Second-Diagonal} = \sum_{i=0}^{n-1} j \sum_{j=0}^{n-1} \left(0.5\,G(i,j)|i-j|\right) \tag{13}$$

$$Variation_{coefficien} = \sigma/\mu \tag{14}$$

$$Entropy_{sum} = -\sum_{i=2}^{2n} G_{x+y}(i)\,Ln\,[G_{x+y}(i)] \tag{15}$$

$$Variance_{sum} = -\sum_{i=2}^{2n} G_{x+y}(i)(i-(\sigma/\mu))^2 \tag{16}$$

$$Average_{sum} = \sum_{i=0}^{n-1} iG_{x+y}(i) \tag{17}$$

$$Entropy_{difference} = -\sum_{i=0}^{n-1} G_{x+y}(i)\,Ln[G_{x+y}(i)] \tag{18}$$

$$Variance_{difference} = -\sum_{i=0}^{n-1} G_{x+y}(i)\,(i-Entropy_{difference})^2 \tag{19}$$

$$Cluster_{\text{shade}} = -\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} G(i,j)[(i+j)-(\mu_x+\mu_y)]^3 \tag{20}$$

$$Cluster_{\text{prominence}} = -\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} G(i,j)[(i+j)-(\mu_x+\mu_y)]^4 \tag{20}$$

$$Entropy1_{x,y} = -\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} G(i,j)\,Ln[G_x(i),G_y(j)],$$

$$Entropy2_{x,y} = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} G_x(i)G_y(j)\,Ln[G_x(i)G_y(j)],$$

$$G_x(i) = \sum_{j=0}^{n-1} G(i,j), G_y(j) = \sum_{i=0}^{n-1} G(i,j),$$

$$Mean_{correlation1} = \frac{Entropy - Entropy1_{x,y}}{\max(Entropy_x, Entropy_y)} \tag{21}$$

$$Mean_{correlation2} = \sqrt{1-\exp^{-2(Entropy2_{x,y}-Entropy)}} \quad (22)$$

There are several difficulties in the GLCM extraction [5][12]: (1) The dimension of the GLCM is directly related to its computational drawbacks for features calculation. (2) There is no pre-defined method for selection of the displacement vector and calculating co-occurrence matrices for different values is computationally cost. For a given image, a large number of features can be computed from GLCM. So, a feature selection method must be used to select the most relevant features. Regarding of extracted features, these features are usually either relevant, redundant or irrelevant. The irrelevant feature does not contribute to the learning process and redundant does not add any additional information to the procedure. Redundant features unnecessarily increase the dimensionality of the feature space and are not expected to improve the classification quality. Whereas the relevant features lead to the best performance. So that feature selection is one of the important steps in order to select best features that give a reduced feature set eventually results in high classification accuracy and also improves the efficacy of training dataset [14][23]. Feature selection is inherently a multi-objective problem with two main objectives of minimizing both the number of features and classification error. In our work, we extracted 22 features from the printed documents of printer dataset. The printer dataset comprises of 1000 images of 10 species of printers. Thus the dimension of the dataset is 1000 x 22. High dimensional feature set could pose a great threat to pattern or image recognition systems. As such, a GA-based feature selection will be used to reduce the number of features needed by the KNN Classifier. A feature subset selection is a map from an *m*-dimensional feature space (input space) to *n*-dimensional feature space (output) [24]. GA is an optimization and search technique based on the principles of genetics and natural selection [14]. The five important issues in the GA are chromosome encoding, fitness evaluation, selection mechanisms, genetic operators and criteria to stop the GA [25]. An initial population is created randomly and evaluated using a fitness function. For binary chromosome employed in this work, a gene value '1' depicts that the particular feature indexed by the position of the '1'is selected. If it is '0', the feature is not selected for evaluation of the concerned chromosome. Herein, the tournament selection mechanism is used due to its simplicity, speed, efficiency, and enforces higher selection pressures on the GA (resulting in higher rate of convergence) and makes sure the worst individual does not get into the next generation [24]. In the tournament selection of size 2, two chromosomes are selected from the population and the better of the two chromosomes using fitness ranking is selected. Tournament selection is performed iteratively until the new population is filled up. Crossover and mutation then form the new population (new generation). The idea behind crossover is that the new chromosome may be better than both of the parents if it takes the best characteristics from each of the parents. On the other hand, after the crossover is performed, mutation takes place. This is to prevent falling all solutions in the population into a local optimum of the solved problem. The mutation changes randomly the new offspring. For binary encoding, a few randomly chosen bits are changed from 1 to 0 or 0 to 1[26]. The fitness of the chromosomes is evaluated using a function commonly referred to as objective function or fitness function [25]. Unlike traditional gradient-based methods, GA's can be used to evolve systems with any kind of fitness measurement functions including those that are non-differentiable, discontinuous. Finding a good fitness measurement can make it easier for GA to evolve a useful system [27]. For a GA to select a subset of features, a fitness function must be defined to evaluate the discriminative capability of each subset of features. In our case, the fitness of each chromosome in the population is evaluated using KNN-based classification error and the cardinality of the selected calculated as [5][22].

$$Fitness = \frac{\alpha}{N_f} + \exp\left(-\frac{1}{N_f}\right) \quad (23)$$

in which α represents the KNN-based classification error, and N_f symbols the cardinality of the selected features. The main objective is to achieve the balance between the classification error minimization with a minimum set of features. As the GA iterates, the individuals (combinatorial set of features) in the current population are evaluated, and their fitness is ranked. Individuals with

lower fitness have a better chance of surviving into the next generation or mating pool. The iterations involved in running the GA ensures that the GA reduce the error rate and picks the individual with the least (best) fitness value since error rate is reported for each chromosome involved and the smallest of error rate is finally picked up by the GA[24]. The adopted GA configuration parameters are shown in Table 1.

In this, the feature's selection procedure is a process of selecting the optimal features that relies on removing the redundant or unnecessary features from the subset guided by the objective function. After obtaining 22 features, the system utilizes GA-based feature selector using a fitness function that integrates both of accuracy (minimize error classification) and feature reduction (minimize the cardinality of the selected features) to aggregate the feature subsets. Based on these optimal features, the testing time can be reduced and the learned classifiers can be simplified. In the final feature subset, the algorithm will select the optimal features from the traditional 22 features in order to get the highest identification rate.

Table 1. GA parameters configuration

GA Parameter	Value
Population size	100
Population type	bit strings
Fitness function	KNN classification error & No. of the selected features
Number of generations	200
Crossover probability	0.8
Mutation probability	0.2
Selection scheme	Tournament of size 2
Elite count	2

Nearest neighbor search is one of the most supervised popular learning and classification techniques that has been proved to be a simple, powerful recognition algorithm, and it is a learning algorithm [28-30]. The KNN is an instance-based classifier that works on the assumption that classification of unknown instances can be identified by relating the unknown to the known instances according to some distance or similarity measure [26]. Given a set of optimal features for each letter "و" in the document, the suggested method employs a 3-Nearest-Neighbor (3NN) classifier. The 3NN classifier is trained with 1000 known feature vectors. The training set is made up of 100 feature vectors from each of the 10 printers listed in Table 2. Each of these feature vectors is independent of each other. To classify an unknown feature vector *X*, the Euclidean distances between *X* and all the known feature vectors are obtained. A majority vote among the 3 smallest distances provides the classification result.

III. Result

In order to test the efficiency and validity of the proposed system, the system prototype was implemented in a modular fashion using MATLAB language release R2015b and was ran and tested using a TOSHIBA PC machine with the following features: Intel (R) Core (TM) i3-2350M CPU @ 2.30GHz, and 4.00 GB of RAM, 64-bit Windows 7 ultimate. In this work, we have used 10 different printers of diverse brands with numerous model and serial numbers are shown in Table 2 that are widely used in the digital evidence laboratory. The first step is to scan the document at 1200 dpi with 8 bits/pixel. Next, the Arabic character "و" in the document (12 point size in Time New Roman font) is extracted in a separated image. The training set consists of 1000 different "و" images for different printers, whereas supplementary 100 images, randomly taken from the same document data set, are used for testing during the identification of the printer source mode. For the evaluation of classification results, the accuracy was chosen as a metric [31][32]:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{24}$$

in which, true positives (*TP*) stands for the number of correctly classified samples, false positives (*FP*) defines the number of wrongly classified samples, true negatives (*TN*) represents the number of

correctly rejected samples and false negatives (*FN*) is the number wrongly rejected samples. For evaluation of the classification per class, precision and recall measures were used: precision is the proportion of positive predictions that are correct, and recall is the proportion of positive samples that are correctly predicted positive [14] [32-34].

$$precision = \frac{TP}{TP + FP} \quad , \quad Recall = \frac{TP}{TP + FN} \tag{25}$$

Table 2. Printer Dataset

	Brand	Model No.	Serial NO.
P_1	Nashua Tec	Spc 410 Aficio	Q7088600729
P_2	Hp	Laser 1102	Vnc4841849
P_3	Samsung	M 332x	Zdfbbjag30006mw
P_4	Samsung	M332x	Zdfbbjcg300023e
P_5	Hp	Laser Jet 1018	Cncig74912
P_6	Canon	LBP3010B	MXBA909688
P_7	Samsung	M 332x	Zdfbbjag300008e
P_8	HP	Laser 1100	CED96852
P_9	Canon	Sansys –Lbp 6020B	Mtma272571
P_{10}	Ricoh	MP 3350 Aficio	FRHRO43547

In a typical forensic printer identification scenario, the accuracy rate is the critical factor to decide the effectiveness of the approach. So, the first experiment tests the classification accuracy for each printer using the optimal features set (approximately from 5 to 7 features instead of the preliminary 22 features depending on the number of samples). The adaptive feature selection algorithm is implemented in this study in order to find the most important features that help to reduce the total evaluation time without the loss of accuracy [18]. The detailed confusion matrix is shown in Table 3. Diagonal element shows the correct classification and the rest shows the incorrect classification. Furthermore, Table 4 depicts the detailed confusion matrix for identification with full features.

Table 3. A confusion matrix of identification results using optimal feature selection.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	Recall
P1	**10**	0	0	0	0	0	0	0	0	0	**100**
P2	0	**10**	0	0	0	0	0	0	0	0	**100**
P3	0	0	**8**	1	0	0	0	0	0	0	**80**
P4	0	0	1	**9**	0	0	0	0	0	0	**90**
P5	0	0	0	0	**8**	2	0	0	0	0	**80**
P6	0	0	0	0	2	**8**	0	0	0	0	**80**
P7	0	0	0	0	0	0	**10**	0	0	0	**100**
P8	0	0	0	0	0	0	0	**10**	0	0	**100**
P9	0	0	0	0	0	0	0	0	**9**	1	**90**
P10	0	0	1	0	0	0	0	0	1	**9**	**90**
Precision	**100**	**100**	**88.8**	**90**	**80**	**80**	**100**	**100**	**90**	**90**	

Table 4. A confusion matrix of identification results using full features.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Recall
P1	10	0	0	0	0	0	0	0	0	0	100
P2	0	10	0	0	0	0	0	0	0	0	100
P3	0	0	8	2	0	0	0	0	0	0	80
P4	0	0	1	6	3	0	0	0	0	0	60
P5	0	0	0	2	7	0	1	0	0	0	70
P6	0	0	0	0	0	10	0	0	0	0	100
P7	0	0	0	0	0	0	8	2	0	0	80
P8	0	0	1	0	0	0	1	8	0	0	80
P9	0	0	0	0	0	0	0	0	9	1	90
P10	0	0	0	0	0	0	0	0	1	9	90
Precision	100	100	85.7	69.2	70	76.9	80	83.3	90	90	

It can be observed from these tables that the suggested system has better recall and precision in many printers, especially for printers numbered P_3 to P_8 as some of these printers differ only in the model number and the serial number. This shows how the proposed system is effective in terms of extracting a set of optimal features that have the ability to accurately distinguish the exact texture features of the printer's documents. After many experiments, the set of optimal features that achieve together higher accuracy are contrast, similarity, mean, diagonal moment, and sum of variance.

The second set of experiments was performed to compare the identification accuracy of the proposed system that employs GA to determine the optimal features and the re-implemented printer identification system introduced in [12] using the same data sets. The results of the presented study revealed that (see Table 5) the use of the 5 optimal features [f_2,f_5,f_7,f_{11},f_{15}] with 3NN classifier generates a further identification rate improvement of 4% for the same method without feature selection phase (22 feature with the same classifier), and 15% improvement compared with the method that relies on GLCM features and 5NN classifier. The performance improvement comes from the correct identification of printers because of using GA to extract optimal features (discriminative features) with the help of the multi-objective fitness function that mixes both of the recognition error and cardinality of the selected features.

Table 5. The identification accuracy rates among different algorithms.

Method	Accuracy rate (%)
Suggested method with optimal features	**91**
Suggested method with full features (3NN)	87
Traditional method using GLCM and 5NN[12]	77.1

The third set of experiments was performed to show how the identification rate of the proposed system depends on the number of samples per printer because if the printer has more enrolled samples, the chance of correct hit increases. The maximum allowed limit of sample documents is 100 per printer and through which they appear different operations on its image, such as font type, font size, rotation, resolution change, and resizing. If the number of samples is above 100 then the returns in performance are however diminishing for every extra sample due to the increase of intra-class printer's variability. In Table 6, as expected, the identification rate increases as the number of samples grows as a result of the increase in inter-class printer's variability. Accuracy rate grows approximately by 2–5% for each increase by 100 of the number of samples in the dataset after 400 samples.

Table 6. Relationship between accuracy rate and the number of samples per printer.

No. of samples	Accuracy (%)
400	82
500	86
600	87
700	88
1000	91

To confirm that the selection of the character "Woo-و" is the most appropriate letter among the group of letters in the Arabic language for printer identification task, the fourth set of experiments is conducted, and the results are shown in Table 7. In general, the "و" character contains a set of bends and circles through which it can extract a total of unique features that can characterize each printer. Furthermore, Table 8 shows the extent to which the accuracy of printer identification is affected by image resolution. As expected the more resolution the better the accuracy. The precision of the image letter through the resolution improves the extraction of the features that lead to enhancing the accuracy rate. Finally, increasing the number of neighbors in KNN classifier may decrease the identification accuracy as illustrated in Table 5 in addition to the increase the computational cost. Sometimes, the increasing of *K* in the KNN classifier would lead to overfitting of the training phase.

Table 7. Relationship between accuracy rate and letter type.

Letter		Accuracy (%)
Alef "ا"	With GA	75
	Without GA	69
Sad "ص"	With GA	70
	Without GA	69
Ain "ع"	With GA	75
	Without GA	73
Woo "و"	With GA	**91**
	Without GA	87

Table 8. Relationship between accuracy rate and letter resolution.

Resolution	Accuracy (%)
300	73
600	75
1200	**91**

The complexity degree of the system depends mainly on the number of samples (n) and the number of generation within GA. As it is difficult to compute the complexity of the system in an accurate way because it is built using MATLAB that requires calling many nested functions; the computational time is used to measure this complexity degree. In general, for the online stage, the system requires approximately 50ms to identify the suspected printer depending on the configuration of the used machine (using 5 optimal features). For the offline phase, the system needs more times for feature extraction and feature selection phases and this time is within 400 to 900 seconds depending on the numbers of samples. Overall, the complexity is roughly $O(n^2)$, which gives us a chance to discover the opportunity of integrating the system with other tools for an integrated online printer identification mechanism entrenched inside an automated real-time digital forensic system especially fraud and forgery research for printers.

IV. Conclusion

This paper presented an intelligent system for identifying printer source for Arabic cursive language that is suitable for printer's fraud and forgery research in forensic science. The suggested system significantly reduces the GLCM features used by KNN classifier through utilizing GA to select the optimal features set. This optimal set achieves two objectives, one is to minimize the classification error and the other is to reduce the number of employed features. The classification accuracy of KNN is considered as one of the variables inside the fitness function for GA. The KNN classifier is one of the most famous neighborhood classifiers in pattern recognition. The integration between a naïve yet delicate KNN classifier and GA as a simple configurable meta-heuristic search engine for optimal features results in a simple yet accurate identification detector. The experiments revealed that the identification accuracy rate can achieve 91% with optimal features set against 86% for traditional features using the same classifier. Future work will focus on improving the robustness of the method by making it work with all brands of devices and improving its accuracy at a lower resolution.

References

[1] G. Ali, A. Mikkilineni, P. Chiang, J. Aleah, G. Chiu and E. Delp, "Intrinsic and Extrinsic Signatures for Information Hiding and Secure printing with Electrophoto-graphy Devices", Proceeding of the International Conference on Digital Printing Technologies, pp. 511–515, Louisiana, 2003.

[2] S. Suh, J. Allebach, G. Chiu and E. Delp, " Printer Mechanism-Level Data Information Embedding and Extraction for Halftone Documents – New Results", International Journal of Imaging Science and Technology, Vol. 2007, No. 2, pp. 549 - 553, 2007.

[3] A. Gadgil, " A Survey of Various Image Processing Techniques for Identification of Printing Technology in Document Forensic Perspective", International Journal of Engineering Inventions, Vol. 1, No. 12, pp. 20 - 28, Dec. 2012.

[4] P. Chiang, A. Mikkilineni, E. Delp, J. Allebach, and G. Chiu, " Extrinsic Signatures Embedding and Detection in Electrophoto-graphic Halftone Images through Laser Intensity Modulation", Proceedings of the International Conference on Digital Printing Technologies, pp. 432-435, USA, 2006.

[5] R. Yadav, K. Goyal, R. Panwar, and N. Khanna, "Comparison of GLCM and IQM for Printer Identification using Printed Documents", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, pp. 6756 – 6760, May 2014.

[6] N. Khannaa, A. Mikkilinenia, A. Martonea, G. Alia, G. Chiub, J. Allebacha and E. Delpa, "A Survey of Forensic Characterization Methods for Physical Devices", International Journal of Digital Forensic, Vol. 3, No. 1, pp. 17- 28, Sep. 2006.

[7] P. Chiang, N. Khanna, A. Mikkilineni, M. Segovia, S. Suh, J. Allebach, G. Chiu, and E. Delp, "Printer and Scanner Forensics", IEEE Signal Processing Magazine, Vol. 26, No. 2, pp.72–83, 2009.

[8] A. Mikkilineni, P. Chiang, G. Ali, G. Chiu, J. Allebach, and E. Delp, "Printer Identification based on Textural Features", Proceedings of the International Conference on Digital Printing Technologies, pp. 306–311, USA, 2005.

[9] N. Khanna, A. Mikkilineni, P. Chiang, M. Ortiz, S. Suh, G. Chiu, J. Allebach, and E. Delp, " Sensor Forensics: Printers, Cameras and Scanners, They Never Lie", Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 20- 23, China, 2-5 July 2007.

[10] J. Mace, "Printer Identification Techniques and Their Privacy Implications", Technical Report, CS-TR-1211, University of Newcastle upon Tyne, UK, pp. 1-11, 2010.

[11] N. Khanna, A. K. Mikkilineni, P. Chiang , M. Ortiz, V. Shah, S. Suh, G. Chiu, J. Allebach, and E. J. Delp, "Printer and Sensor Forensics", Proceedings of the IEEE International Conference on Signal Processing Applications for Public Security and Forensics, pp. 1- 8, USA, April 2007.

[12] A. Mikkilineni, P. Chiang, G. Ali, G. Chiu, J. Allebach and E. Delp, " Printer Identification based on Gray level Co-occurrence Features for Security and Forensic Applications", Proceedings of the International Conference on Security, Steganography and Watermarking of Multimedia Contents, pp. 430 – 440, USA, 2005.

[13] J. Choi, H. Lee, and K. Lee, "Color Laser Printer Forensic based on Noisy Feature and Support Vector Machine Classifier ", International Journal of Multimedia Tools and Application, Vol. 67, Issue 2, pp. 363 - 382, Nov. 2013.

[14] G. Kumar, G. Ramachandra, and K. Nagamani, "An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 2, pp. 272 – 277, Feb. 2014.

[15] S. M. Saad, "Application of Fuzzy Logic and Genetic Algorithm in Biometric Text-Independent Writer Identification ", IET Information Security, Vol. 5, No. 1, pp. 1-9, March 2011

[16] H. Lee, and J. Choi, "Identifying Color Laser Printer using Noisy Feature and Support Vector Machine", Proceedings of the fifth IEEE International Conference on Ubiquitous Information Technologies and Applications, pp. 1-6, China, 2010.

[17] M. Tsai, J. Yin, I. Yuadi, and J. Liu," Digital Forensics of Printed Source Identification for Chinese Characters", International Journal of Multimedia Tools and Applications, Vol. 73, No. 3, pp. 2129 - 2155, Dec. 2014.

[18] M. Tsai, C. Hsu, J. Yin, and I. Yuad, "Japanese Character based Printed Source Identification", Proceedings of the IEEE International Conference on Circuits and Systems, pp. 2800 - 2803, Taiwan, 24-27 May 2015.

[19] Q. Zhou, Y. Yan, T. Fang, X. Luo, and Q. Chen" Text-Independent Printer Identification based on Texture Synthesis", International Journal on Multimedia Tools and Applications, Vol. 75, No. 10, pp. 5557-5580. 2016.

[20] M. Tsai, J. Liu, C. Wang, and C. Chuang, "Source Color Laser Printer Identification using Discrete Wavelet Transform and Feature Selection Algorithms", Proceedings of the IEEE International Symposium on Circuits and Systems, pp. 2633 - 2636, Brazil, May 2011.

[21] M. Tsai, and J. Liu, "Digital Forensics for Printed Source Identification", Proceedings of the IEEE International Conference on Circuits and Systems, pp. 2347–2350, Taiwan, May 2013.

[22] O. Abouelatta," Classification of Copper Alloys Microstructure using Image Processing and Neural Network", Journal of American Science, Vol. 9, No. 6, pp.213-223, 2013.

[23] R. Jain " Application of KNN-Genetic Algorithm for Analyzing Student Learning in Educational Data Mining Paradigm" International Journal of Innovative Research in Computer and Communication Engineering , Vol. 4, Issue 6, pp. 10319 - 10323, June 2016.

[24] B. Oluleye, A. Leisa, J. Leng, and D. Dean "A Genetic Algorithm - based Feature Selection." British Journal of Mathematics & Computer Science, Vol. 4. No. 21, pp. 889-905, 2014.

[25] B. Oluleye, A. Leisa, J. Leng, and D. Dean " Zernike Moments and Genetic Algorithm: Tutorial and Application." British Journal of Mathematics & Computer Science, Vol. 4. No. 15, pp. 2217-2236, 2014.

[26] C. Gunavathi, and K. Premalatha. "Performance Analysis of Genetic Algorithm with KNN and SVM for Feature Selection in Tumor Classification." International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol. 8, No. 8, pp. 1490-1497, 2014.

[27] B. Jayasekara, A. Jayasiri, L. Udawatta "An Evolving Signature Recognition System", Proceedings of the IEEE International Conference on Industrial and Information Systems. pp. 529—534, Sri Lanka, 2006.

[28] N. Suguna, and K. Thanushkodi. "An Improved K-Nearest Neighbor Classification using Genetic Algorithm", International Journal of Computer Science Issues, Vol. 7, No 2, pp. 18-21, July 2010.

[29] A. Mikkilineni, O. Arslan, P. Chiang, R. Kumontoy, J. Allebach, G. Chiu, and E. Delp, " Printer Forensics using SVM Techniques", Proceedings of the International Conference on Digital Printing Technologies, pp. 223 – 226, Maryland, 2005.

[30] A. Mikkilineni, N. Khanna, and E. Delp, " Texture based Attacks on Intrinsic Signature based Printer Identification", Proceedings of the International Conference on Media Forensic and Security, pp. 1-12, California, Jan. 2010.

[31] M. Saraswat, K. Goswami, and A. Tiwari," Object Recognition using Texture based Analysis", International Journal of Computer Science and Information Technologies, Vol. 4, No. 6, pp. 775-782, 2013.

[32] S. Elkasrawi, and F. Shafait, " Printer Identification using Supervised Learning for Document Forgery Detection", Proceedings of the 11th IEEE International Workshop on Document Analysis Systems, pp. 146 – 150, France, 7-10 April 2014.

[33] Y. Wu, X. Kong, X. You, and Y. Guo, "Printer Forensics based on Page Document's Geometric Distortion", Proceedings of the IEEE International Conference on Image Processing, pp. 2909 – 2912, Egypt, 7-10 Nov. 2009.

[34] W. Deng, Q. Chen, F. Yuan, and Y. Yan, "Printer Identification based on Distance Transform", Proceedings of The IEEE International Conference on Intelligent Networks and Intelligent Systems, pp.565-568. China, 2008.