Selecting Significant Features for Authorship Invarianceness in Writer Identification

Azah Kamilah Muda, Satrya Fajri Pratama, Yun-Huoy Choo, and Noor Azilah Muda

Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka. Hang Tuah Jaya,
76100 Durian Tunggal, Melaka, Malaysia
azah@utem.edu.my, rascove@yahoo.com, huoy@utem.edu.my, azilah@utem.edu.my,

Abstract. Handwriting is individualistic. The uniqueness of shape and style of handwriting can be used to identify the significant features in authenticating the author of writing. Acquiring these significant features leads to an important research in Writer Identification domain where to find the unique features of individual which also known as Individuality of Handwriting. It relates to invarianceness of authorship where invarianceness between features for intra-class (same writer) is lower than inter-class (different writer). This paper discusses and reports the exploration of significant features for invarianceness of authorship from global shape features by using feature selection technique. The promising results show that the proposed method is worth to receive further exploration in identifying the handwritten authorship.

Keywords: feature selection, authorship invarianceness, significant features.

1 Introduction

Feature selection has become the focus of research area for a long time. The purpose of feature selection is to obtain the most minimal sized subset of features [1]. Practical experience has shown that if there is too much irrelevant and redundant information present, the performance of a classifier might be degraded. Removing these irrelevant and redundant features can improve the classification accuracy.

The three popular methods of feature selection are filter method, wrapper method, and embedded method has been presented in [2]. Filter method assesses the relevance of features [3], wrapper method uses an induction algorithm [4], while embedded method do the selection process inside the induction algorithm [5]. Studies have shown that there are no techniques more superior compared to others [6].

Writer Identification (WI) can be included as a particular kind of dynamic biometric in pattern recognition for forensic application. WI distinguishes writers based on the shape or individual style of writing while ignoring the meaning of the word or character written. The shape and style of writing are different from one person to another. Even for one person, they are different in times. However,

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 637–651, 2011.

everyone has their own style of writing and it is individualistic. It must be unique feature that can be generalized as significant individual features through the handwriting shape.

Many previous works on WI problem has been tried to be solved based on the image processing and pattern recognition technique [7], [8], [9], [10], [11] and involved feature extraction task. Many approaches have been proposed to extract the features for WI. Mostly, features are extracted from the handwriting focus on rigid characteristics of the shape such as [7], [9], [11], [12], [13], [14], [15], [16], [17] except by [18] and [19], focus on global features.

The main issue in WI is how to acquire the features that reflect the author of handwriting. Thus, it is an open question whether the extracted features are optimal or near-optimal to identify the author. Extracted features may include many garbage features. Such features are not only useless in classification, but sometimes degrade the performance of a classifier designed on a basis of a finite number of training samples [20]. The features may not be independent of each other or even redundant. Moreover, there may be features that do not provide any useful information for the task of writer identification [21]. Therefore, feature extraction and selection of the significant features are very important in order to identify the writer, moreover to improve the classification accuracy.

Thus, this paper focuses on identifying the significant features of word shape by using the proposed feature selection technique prior the identification task. The remainder of the paper is structured as follows. In next section, an overview of individuality of handwriting is given. Global feature representation by United Moment Invariant is described in Section 3. Section 4 provides an overview of feature selection techniques, followed by the proposed approach to identify the significant features in Section 5. Finally, conclusion and future work is drawn in Section 6.

2 Authorship Invarianceness

Handwriting is individual to personal. Handwriting has long been considered individualistic and writer individuality rests on the hypothesis that each individual has consistent handwriting [10], [18], [23], [24], [25]. The relation of character, shape and the style of writing are different from one to another.

Handwriting analysis consists of two categories, which are handwriting recognition and handwriting identification. Handwriting recognition deals with the contents conveyed by the handwritten word, while handwriting identification tries to differentiate handwritings to determine the author. There are two tasks in identifying the writer of handwriting, namely identification and verification. Identification task determines the writer of handwriting from many known writers, while verification task determines whether one document and another is written by the same writer.

The challenge in WI is how to acquire the features that reflect the author for these variety styles of handwriting [7], [9], [12], [13], [15], [24], either for one writer or many writers. These features are required to classify in order to identify the variance between features for same writer is lower than different writer which known as Authorship Invarianceness. Among these features are exists the significant individual features which directly unique to those individual. Figure 1 shows that each person has its individuality styles of writing. The shape is slightly different for the same writer and quite difference for different writers.

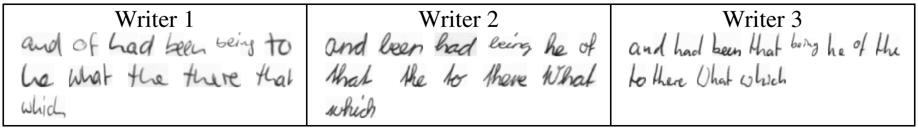


Fig. 1. Different Word for Different Writer

3 Global Features Representation

In pattern recognition problem, there are many shape representations or description techniques have been explored in order to extract the features from the image. Generally it can be classified into two different approaches when dealing with handwritten word problem, which are analytic (local / structural approach) and holistic (global approach) [26], [27]. For the each approach, it is divided into two method, which are region-based (whole region shape) methods and contour-based (contour only) methods. Holistic approach represent shape as a whole, meanwhile analytic approach represents image in sections. In this work, holistic approach of United Moment Invariant (UMI) is chosen due to the requirement of cursive word is needed to extract as one single indivisible entity. This moment function of UMI is applied in feature extraction task.

The choice of using holistic approach is not only based on the holistic advantages, but also due to its capability of using word in showing the individuality for writer identification problem as mentioned in [18] and holds immense promise for realizing near-human performance [28]. The holistic features and matching schemes must be coarse enough to be stable across exemplars of the same class such as a variety of writing styles [29]. This is aligning with this work where to extract the unique global features from word shape in order to identify the writer.

Global features extracted with this holistic approach are invariant with respect to all different writing styles [29]. Words in general may be cursive, minor touching discrete, purely discrete, one or two characters are isolated and others are discrete or mixture of these style and it still as one word. Global technique in holistic approach will extract all of these styles for one word as one whole shape. Shape is an important representation of visual image of an object. It is a very powerful feature when it is used in similarity search. Unlike color and texture features, the shape of an object is strongly tied to the object functionality and identity [30]. Furthermore, the use of holistic approach is shown to be very effective in lexicon reduction [31], moreover to increase the accuracy of classification.

3.1 United Moment Invariant Function

Moment Function has been used in diverse fields ranging from mechanics and statistics to pattern recognition and image understanding [32]. The use of moments in image analysis and pattern recognition was inspired by [33] and [34]. [33] first presented a set of seven-tuplet moments that invariant to position, size, and orientation of the image shape. However, there are many research have been done to prove that there were some drawback in the original work by [33] in terms of invariant such as [35], [36], [37], [38], [39], and [40]. All of these researchers proposed their method of moment and tested on feature extraction phase to represents the image. A good shape descriptor should be able to find perceptually similar shape where it is usually means rotated, translated, scaled and affined transformed shapes. Furthermore, it can tolerate with human beings in comparing the image shapes. Therefore, [41] derived United Moment Invariants (UMI) based on basic scaling transformation by [33] that can be applied in all conditions with a good set of discriminate shapes features. Moreover, UMI never been tested in WI domain. With the capability of UMI as a good description of image shape, this work is explored its capability of image representation in WI domain.

[41] proposed UMI with mathematically related to GMI by [33] by considering (1) as normalized central moments:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\frac{p+q+2}{2}}}, \quad p+q = 2, 3, \ldots . \tag{1}$$

and (2) in discrete form. Central and normalized central moments are given as:

$$\mu'_{pq} = \rho^{p+q}\mu_{pq}, \qquad \eta'_{pq} = \rho^{p+q}\eta_{pq} = \frac{\rho^{p+q}}{\mu_{00}^{\frac{p+q+2}{2}}}\mu_{pq} . \tag{2}$$

and improved moment invariant by [43] is given as:

$$\eta'_{pq} = \frac{\mu_{pq}}{\mu_{00}^{p+q+1}}. \tag{3}$$

(1) to (3) have the factor μ_{pq}. Eight feature vector derived by [40] are listed below:

$$\theta_1 = \frac{\sqrt{\phi_2}}{\phi_1}. \qquad \theta_2 = \frac{\phi_6}{\phi_1\phi_4}.$$
$$\theta_3 = \frac{\sqrt{\phi_5}}{\phi_4}. \qquad \theta_4 = \frac{\phi_5}{\phi_3\phi_4}.$$
$$\theta_5 = \frac{\phi_1\phi_6}{\phi_2\phi_3}. \qquad \theta_6 = \left(\phi_1 + \sqrt{\phi_2}\right)\frac{\phi_3}{\phi_6}.$$
$$\theta_7 = \frac{\phi_1\phi_5}{\phi_3\phi_6}. \qquad \theta_8 = \frac{\phi_3+\phi_4}{\sqrt{\phi_5}}. \tag{4}$$

where ϕ_i are Hu's moment invariants.

4 Feature Selection

Feature selection has become an active research area for decades, and has been proven in both theory and practice [44]. The main objective of feature selection is to select the minimally sized subset of features as long as the classification accuracy does not significantly decreased and the result of the selected features class distribution is as close as possible to original class distribution [1]. In contrast to other dimensionality reduction methods like those based on projection or compression, feature selection methods do not alter the original representation of the variables, but merely select a subset of them. Thus, they preserve the original semantics of the variables. However, the advantages of feature selection methods come at a certain price, as the search for a subset of relevant features introduces an additional layer of complexity in the modeling task [2]. In this work, feature selection is explored in order to find the most significant features which by is the unique features of individual's writing. The unique features a mainly contribute to the concept of Authorship Invarianceness in WI.

There are three general methods of feature selection which are filter method, wrapper method, and embedded method [45]. Filter method assesses the relevance of features by looking only at the intrinsic properties of the data. A feature relevance score is calculated, and low-scoring features are removed [3]. Simultaneously, wrapper method uses an induction algorithm to estimate the merit of feature subsets. It explores the space of features subsets to optimize the induction algorithm that uses the subset for classification [4]. On the other hand, in embedded method, the selection process is done inside the induction algorithm itself, being far less computationally intensive compared with wrapper methods [5]. However, the focus of this paper is to explore the use of wrapper methods. Wrapper strategies for feature selection use an induction algorithm to estimate the merit of feature subsets. The rationale for wrapper methods is that the induction method that will ultimately use the feature subset should provide a better estimate of accuracy than a separate measure that has an entirely different inductive bias [3].

The wrapper method is computationally demanding, but often is more accurate. A wrapper algorithm explores the space of features subsets to optimize the induction algorithm that uses the subset for classification. These methods based on penalization face a combinatorial challenge when the set of variables has no specific order and when the search must be done over its subsets since many problems related to feature extraction have been shown to be NP-hard [4]. Advantages of wrapper methods include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. A common drawback of these methods is that they have a higher risk of over-fitting than filter methods and are very computationally intensive, especially if building the classifier has a high computational cost [2]. There are several wrapper techniques, however only two techniques will be discussed here. These techniques are Sequential Forward Selection and Sequential Forward Floating Selection. Figure 2 depicts wrapper feature selection method.

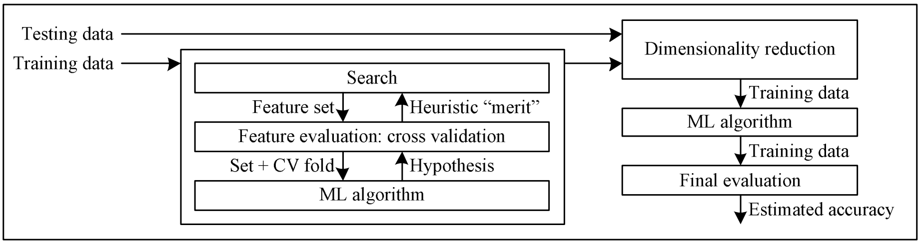


Fig. 2. Wrapper Feature Selection [3]

Sequential Forward Selection (SFS) is introduced by [46] which proposed the best subset of features Y_0 that is initialized as the empty set. The feature x^+ that gives the highest correct classification rate $J(Y_k + x^+)$ is added to Y_k at the each step along with the features which already included in Y_k. The process continues until the correct classification rate given by Y_k and each of the features not yet selected does not increase. SFS performs best when the optimal subset has a small number of features. When the search is near the empty set, a large number of states can be potentially evaluated, and towards the full set, the region examined by SFS is narrower since most of the features have already been selected. The algorithm of SFS is shown as below:

1. Start with the empty set $Y_0 = \{\emptyset\}$
2. Select the next best feature $x^+ = \text{argmax}_{x^+ \notin Y_k}[J(Y_k + x^+)]$
3. Update $Y_{k+1} = Y_k + x^+; k = k + 1$
4. Go to step 2

However, this method suffers from the nesting effect. This means that a feature that is included in some step of the iterative process cannot be excluded in a later step. Thus, the results are sub-optimal. Therefore, the Sequential Forward Floating Selection (SFFS) method was introduced by [47] to deal with the nesting problem. In SFFS, Y_0 is initialized as the empty set and in each step a new subset is generated first by adding a feature x^+, but after that features x^- is searched for to be eliminated from Y_k until the correct classification rate $J(Y_k - x^-)$ decreases. The iterations continue until no new variable can be added because the recognition rate $J(Y_k + x^+)$ does not increase. The algorithm is as below.

1. Start with the empty set $Y_0 = \{\emptyset\}$
2. Select the next best feature $x^+ = \text{argmax}_{x^+ \notin Y_k}[J(Y_k + x^+)]$
3. Update $Y_{k+1} = Y_k + x^+; k = k + 1$
4. Remove the worst feature $x^- = \text{argmax}_{x^- \in Y_k}[J(Y_k - x^-)]$
5. If $J(Y_k - x^-) > J(Y_k)$
   Update $Y_{k+1} = Y_k - x^-; k = k + 1$
   Go to 3
   Else
   Go to 2

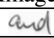
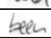
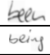
5 Proposed Approach

The experiments described in this paper are executed using the IAM database [48]. Various types of word images from IAM database are extracted using UMI to represent the image into feature vector. The selection of significant features using the wrapper methods are performed prior the identification task. The selected features which produce highest accuracy from the identification task are identified as the optimal significant features for WI in this work and also known as unique features of individual's writing.

5.1 Extracting Features

Feature extraction is a process of converting input object into feature vectors. The extracted features are in real value and unique for each word. A set of moments computed from digital image using UMI represents global characteristics of an image shape, and provides a lot of information about different types of geometrical features of the image [49]. Different types of words from IAM database such as 'the', 'and', 'where' and others have been extracted from one author. In this paper, a total number of 4400 instances are extracted to be used for the experiments, and are randomly divided into five different datasets to form training and testing dataset. Table 1 is the example of feature invariant of words using UMI with eight features vector for the each image.

Table 1. Example of Feature Invariant

Image	Feature 1	………	Feature 7	Feature 8
	0.217716	………..	1.56976	1.82758
	0.115774	………..	0.0552545	0.499824
	0.369492	…………	0.124305	0.580407

Extracted features can be divided into micro and macro feature classes which are local and global features. Local features denote the constituent parts of objects and the relationships, meanwhile global features describing properties of the whole object [50]. Good features are those satisfying two requirements which are small intra-class invariance and large inter-class invariance [51]. This can be defined as invarianceness of authorship in WI.

Invarianceness of authorship in WI shows the similarity error for intra-class (same-writer) is small compared to inter-class (different-writers) for the same words or different words. This is due to the individual features of handwriting's style which has been proof in many researchers such as [18], [24], and [52]. Related to this paper, the objective is to make contributions towards this scientific validation using the proposed techniques for selecting the significant features in order to proof the authorship of invarianceness in WI. The uniqueness of this work is to find the significant feature which actually is the unique features of individual's writing. The invarianceness of authorship relates to individuality of handwriting with the unique features of individual's writing. The highest accuracy of selected features proofs the invarianceness of authorship for intra-class is lower than inter-class where each individual's writing contains the unique styles of handwriting that is different with other individual. To achieve this, the process of selecting significant features is carried out using the proposed wrapper method before identification task.

5.2 Selecting Significant Features

Two commonly used wrapper method discussed earlier will be used to determine the significant features. These feature selection techniques will be using Modified Immune Classifier (MIC) [53] as their classifier. Every experiment has been performed using ten-fold cross-validation. These feature selection techniques will be executed five times to ensure the performance is stable and accurate.

In order to justify the quality of feature subset produced by each method, other state-of-the-art feature selection techniques are also used, which are Correlation-based Feature Selection (CFS) [3], Consistency-based Feature Selection, also known as Las Vegas Filter (LVF) [54], and Fast Correlation-based Filter (FCBF) [55]. Other classifiers are also being used for SFS to further validate the result, which are Naïve Bayes [56] and Random Forest [57] classifier. These feature selection techniques are provided in WEKA [58]. Justification of these feature selection techniques has been presented in [22]. Table 2 is the result of selection for each feature invariant data set.

Table 2. Experimental Results on Feature Selection

Method	Execution	Set A	Set B	Set C	Set D	Set E	Intersection
SFS (using MIC)	Execution #1	f2, f3, f6, f8	f2, f3, f4, f6, f8	f1, f3, f6, f7, f8	f3, f6, f8	f1, f2, f3, f5, f6, f7	**f3, f6**
	Execution #2	f1, f3, f4, f6, f8	f1, f3, f4, f5, f6, f8	f1, f3, f4, f6, f8	f1, f2, f3, f6	f1, f3, f6	**f3, f6**
	Execution #3	f2, f3, f4, f5, f6, f8	f1, f3, f6, f7, f8	f1, f3, f6, f8	f2, f3, f6, f7, f8	f3, f4, f5, f6, f7, f8	**f3, f6, f8**
	Execution #4	f2, f3, f6, f8	f1, f3, f4, f5, f6, f8	f1, f2, f3, f4, f5, f6	f1, f3, f4, f5, f6	f1, f3, f6	**f3, f6**
	Execution #5	f3, f6, f7, f8	f1, f2, f3, f6	f2, f3, f4, f5, f6, f7, f8	f1, f3, f6, f8	f2, f3, f6, f8	**f3, f6**
	Intersection	**f3, f6, f8**	**f3, f6**	**f3, f6**	**f3, f6**	**f3, f6**	**f3, f6**
SFFS (using MIC)	Execution #1	f1, f3, f6	f2, f3, f4, f6	f1, f3, f4, f5, f6, f8	f1, f3, f6, f8	f3, f4, f6, f7, f8	**f3, f6**
	Execution #2	f1, f3, f5, f6, f7, f8	f3, f4, f6, f7, f8	f1, f2, f3, f4, f6, f8	f1, f3, f4, f5, f6, f8	f2, f3, f5, f6	**f3, f6**
	Execution #3	f2, f3, f4, f5, f6, f7, f8	f2, f3, f5, f6, f8	f1, f2, f3, f6, f7, f8	f2, f3, f6, f8	f2, f3, f6, f8	**f3, f6, f8**
	Execution #4	f3, f4, f6, f8	f1, f2, f3, f6	f3, f6, f7, f8	f3, f4, f6, f8	f3, f6, f8	**f3, f6**
	Execution #5	f2, f3, f4, f5, f6, f7	f1, f2, f3, f6, f7, f8	f2, f3, f4, f5, f6, f8	f1, f3, f6, f8	f3, f6, f7, f8	**f3, f6**
	Intersection	**f3, f6**	**f3, f6**	**f3, f6, f8**	**f3, f6, f8**	**f3, f6**	**f3, f6**

Table 2. (*continued*)

Method	Execution	Set A	Set B	Set C	Set D	Set E	Intersection
CFS	Execution #1	f1, f2, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	f1, f3, f4, f5, f6, f7	f1, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	**f1, f3, f5, f7**
	Execution #2	f1, f2, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	f1, f3, f4, f5, f6, f7	f1, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	**f1, f3, f5, f7**
	Execution #3	f1, f2, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	f1, f3, f4, f5, f6, f7	f1, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	**f1, f3, f5, f7**
	Execution #4	f1, f2, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	f1, f3, f4, f5, f6, f7	f1, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	**f1, f3, f5, f7**
	Execution #5	f1, f2, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	f1, f3, f4, f5, f6, f7	f1, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	**f1, f3, f5, f7**
	Intersection	**f1, f2, f3, f5, f7, f8**	**f1, f3, f4, f5, f6, f7**	**f1, f3, f4, f5, f6, f7**	**f1, f3, f5, f7, f8**	**f1, f3, f4, f5, f6, f7**	**f1, f3, f5, f7**
LVF	Execution #1	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	**f2, f3, f4, f6**
	Execution #2	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	**f2, f3, f4, f6**
	Execution #3	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	**f2, f3, f4, f6**
	Execution #4	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	**f2, f3, f4, f6**
	Execution #5	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	**f2, f3, f4, f6**
	Intersection	**f2, f3, f4, f6**	**f2, f3, f4, f6**	**f2, f3, f4, f6**	**f2, f3, f4, f6**	**f2, f3, f4, f6**	**f2, f3, f4, f6**
FCBF	Execution #1	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	**f1, f2, f3, f4, f5, f6, f7, f8**
	Execution #2	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	**f1, f2, f3, f4, f5, f6, f7, f8**
	Execution #3	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	**f1, f2, f3, f4, f5, f6, f7, f8**
	Execution #4	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	**f1, f2, f3, f4, f5, f6, f7, f8**
	Execution #5	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	**f1, f2, f3, f4, f5, f6, f7, f8**
	Intersection	**f1, f2, f3, f4, f5, f6, f7, f8**	**f1, f2, f3, f4, f5, f6, f7, f8**	**f1, f2, f3, f4, f5, f6, f7, f8**	**f1, f2, f3, f4, f5, f6, f7, f8**	**f1, f2, f3, f4, f5, f6, f7, f8**	**f1, f2, f3, f4, f5, f6, f7, f8**
SFS (using Naïve Bayes)	Execution #1	f1, f2, f3, f5, f8	f3, f4	f3	f1, f3, f4	f3	**f3**
	Execution #2	f1, f2, f3, f5, f8	f3, f4	f3	f1, f3, f4	f3	**f3**

Table 2. (*continued*)

Method	Execution	Set A	Set B	Set C	Set D	Set E	Intersection
	Execution #3	f1, f2, f3, f5, f8	f3, f4	f3	f1, f3, f4	f3	**f3**
	Execution #4	f1, f2, f3, f5, f8	f3, f4	f3	f1, f3, f4	f3	**f3**
	Execution #5	f1, f2, f3, f5, f8	f3, f4	f3	f1, f3, f4	f3	**f3**
	Intersection	**f1, f2, f3, f5, f8**	**f3, f4**	**f3**	**f1, f3, f4**	**f3**	**f3**
SFS (using Random Forest)	Execution #1	f3	f3	f3	f3	f3	**f3**
	Execution #2	f3	f3	f3	f3	f3	**f3**
	Execution #3	f3	f3	f3	f3	f3	**f3**
	Execution #4	f3	f3	f3	f3	f3	**f3**
	Execution #5	f3	f3	f3	f3	f3	**f3**
	Intersection	**f3**	**f3**	**f3**	**f3**	**f3**	**f3**

Based on the feature selection results, it is shown that these feature selection techniques yield different subset with different size. It is shown that SFS with Naïve Bayes in set C and E and Random Forest in all set select only one feature. These two are not capable to reduce the number of features partially due to the nature of the data itself. It is also known prone to over-fitting to some datasets and cannot handle large numbers of irrelevant features, thus it is not capable to reduce the number of features,

On the other hand, FCBF is shown to unable reduce the number of features, this is because this feature selection technique is more suitable when handling high-dimensional data, because it analyze the correlation between features, which is feature relevancy and feature redundancy. Thus, these methods will perform poorly when they failed to find the correlation between features, or they overestimate the correlation between features. In other domain of pattern recognition, the results obtained from FCBF and SFS with Naïve Bayes and Random Forest can be considered as suboptimal result, however in this WI domain, these feature selection techniques is still considered to achieve the purpose of the experiment. This is because the purpose of feature selection in WI is not only to reduce the number of features; instead it is to determine the most significant features (unique features). Thus, FCBF considers all features are significant, while SFS with Naïve Bayes and Random Forest consider that the selected features are the most significant feature.

On the contrary, the rest of the techniques (SFS and SFFS with MIC, CFS, and LVF) are able to identify the significant features. It is also worth mentioning that although these feature selection techniques yield different features, they seem to always include the third feature (f3) in their results. Therefore, it can be concluded that the third feature (f3) is the most significant feature, and it is chosen as significant unique feature in order to proof the invarianceness of authorship in this work.

5.3 Identifying the Authorship Using Significant Features

The selected significant features from every feature selection techniques must be justified and validated through identification performance. In order to justify the quality of feature subset produced by each method, the feature subsets are tested against classification, which uses MIC as the classifier. Table 3 is the result of identification accuracy for each feature subset.

Table 3. Experimental Results on Identification Accuracy (%)

Method	Execution	Set A	Set B	Set C	Set D	Set E	Average
SFS (using MIC)	Execution #1	97.40	97.18	96.92	96.14	96.94	**96.92**
	Execution #2	97.29	97.77	96.01	96.47	95.80	**96.67**
	Execution #3	97.63	97.30	95.78	96.80	97.05	**96.91**
	Execution #4	97.40	97.77	97.26	96.80	95.80	**97.01**
	Execution #5	97.51	96.59	97.38	96.14	96.49	**96.82**
	Average	**97.45**	**97.32**	**96.67**	**96.47**	**96.42**	**96.87**
SFFS (using MIC)	Execution #1	96.95	96.71	97.04	96.14	96.49	**96.66**
	Execution #2	97.40	97.18	96.58	97.13	96.94	**97.05**
	Execution #3	94.35	97.41	97.04	96.03	96.49	**96.26**
	Execution #4	97.06	96.59	96.58	96.14	96.03	**96.48**
	Execution #5	97.51	97.18	97.04	96.14	96.60	**96.89**
	Average	**96.66**	**97.02**	**96.85**	**96.32**	**96.51**	**96.67**
CFS	Execution #1	94.24	97.18	97.18	94.01	97.18	**95.95**
	Execution #2	94.24	97.18	97.18	94.01	97.18	**95.95**
	Execution #3	94.24	97.18	97.18	94.01	97.18	**95.95**
	Execution #4	94.24	97.18	97.18	94.01	97.18	**95.95**
	Execution #5	94.24	97.18	97.18	94.01	97.18	**95.95**
	Average	**94.24**	**97.18**	**97.18**	**94.01**	**97.18**	**95.95**
LVF	Execution #1	97.40	97.40	97.40	97.40	97.40	**97.40**
	Execution #2	97.40	97.40	97.40	97.40	97.40	**97.40**
	Execution #3	97.40	97.40	97.40	97.40	97.40	**97.40**
	Execution #4	97.40	97.40	97.40	97.40	97.40	**97.40**
	Execution #5	97.40	97.40	97.40	97.40	97.40	**97.40**
	Average	**97.40**	**97.40**	**97.40**	**97.40**	**97.40**	**97.40**
FCBF	Execution #1	97.74	97.74	97.74	97.74	97.74	**97.74**
	Execution #2	97.74	97.74	97.74	97.74	97.74	**97.74**
	Execution #3	97.74	97.74	97.74	97.74	97.74	**97.74**
	Execution #4	97.74	97.74	97.74	97.74	97.74	**97.74**
	Execution #5	97.74	97.74	97.74	97.74	97.74	**97.74**
	Average	**97.74**	**97.74**	**97.74**	**97.74**	**97.74**	**97.74**
SFS (using Naïve Bayes)	Execution #1	93.79	88.47	80.23	92.09	80.23	**86.96**
	Execution #2	93.79	88.47	80.23	92.09	80.23	**86.96**
	Execution #3	93.79	88.47	80.23	92.09	80.23	**86.96**
	Execution #4	93.79	88.47	80.23	92.09	80.23	**86.96**
	Execution #5	93.79	88.47	80.23	92.09	80.23	**86.96**
	Average	**93.79**	**88.47**	**80.23**	**92.09**	**80.23**	**86.96**
SFS (using Random Forest)	Execution #1	80.23	80.23	80.23	80.23	80.23	**80.23**
	Execution #2	80.23	80.23	80.23	80.23	80.23	**80.23**
	Execution #3	80.23	80.23	80.23	80.23	80.23	**80.23**
	Execution #4	80.23	80.23	80.23	80.23	80.23	**80.23**
	Execution #5	80.23	80.23	80.23	80.23	80.23	**80.23**
	Average	**80.23**	**80.23**	**80.23**	**80.23**	**80.23**	**80.23**

Based on the results, the accuracy is at its highest when the number of features is between 4-7 features. It is shown that FCBF produces the best accuracy (97.74%) and equal with the original dataset performance (97.74%). However, the number of features produced by FCBF is equal with the actual set (8 features). Meaning that, FCBF needs all features to produce the best performance. The second best accuracy is produced by LVF (97.40%). The results of LVF are shown to be stable, regardless of dataset and the number of execution. This is because the nature of the data that is consistent allows LVF to perform well.

On the other hand, both SFS with MIC (96.87%) and SFFS with MIC (96.67%) with lower number of features still can obtain almost similar performance, although it is slightly lower than original dataset (97.74%). These feature selection technique outperform some other techniques (CFS, SFS using Naïve Bayes, and SFS using Random Forest). This is due to the behavior of these techniques which can specifically identify the unique features in dataset, therefore it is resulting the highest performance. Besides that, the wrapper technique is able to recognize importance of each feature in every iteration. However, due to the nature of both Naïve Bayes and Random Forest, the performance of SFS is deteriorating (86.96% and 80.23%).

These techniques are both capable to identify the most significant features and at the same time they validate the invarianceness of authorship concept where the invariance between features for intra-class is lower than inter-class. As a normal practice in pattern recognition, it can be achieved by calculating the invariance for intra-class and inter-class using Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - r_i| \ . \qquad (5)$$

The result in Table 4 shows that the invarianceness of authorship is proven where the invarianceness between features using selected features for intra-class (same author) is smaller compared to inter-class (different author). This conforms the significant features is relate to invarianceness of authorship on WI.

Table 4. Identification Accuracy Results (%)

Various words	**1 writer**	**10 writers**	**20 writers**
20 words	0.278666	0.295112	0.524758
40 words	0.289052	0.295236	0.512279
60 words	0.282408	0.293509	0.527289
80 words	0.270236	0.3018	0.520221
100 words	0.281886	0.355219	0.544051

It is also shown that CFS is also capable to obtain good result (95.95%), although it is not as good as LVF, SFS and SFFS with MIC. Although FCBF is the enhancement of CFS, it is shown that CFS is still better than FCBF in some dataset. This is because FCBF determines the correlation between features faster than CFS, which may causing the technique to overestimate the correlation between features, thus causing it to select all the features.

6 Conclusion and Future Work

The exploration of significant unique features relates to authorship invarianceness has been presented in this paper. A scientific validation has been provided as evidence of significant features can be used to proof the authorship invarianceness in WI. In future works, the selected unique features will be further explored with other classifier to confirm these features can be used as optimized features with higher accuracy. An improved sequential forward selection will also be developed to better adapt the nature of the data, and thus increase the performance.

References

1. Dash, M., Liu, H.: Feature Selection for Classification. J. Intelligent Data Analysis 1, 131–156 (1997)
2. Saeys, Y., Inza, I., Larranaga, P.: A Review of Feature Selection Techniques in Bioinformatics. J. Bioinformatics 23, 2507–2517 (2007)
3. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. PhD Thesis, University of Waikato (1999)
4. Gadat, S., Younes, L.: A Stochastic Algorithm for Feature Selection in Pattern Recognition. J. Machine Learning Research 8, 509–547 (2007)
5. Portinale, L., Saitta, L.: Feature Selection: State of the Art. In: Feature Selection, pp. 1–22. Universita del Piemonte Orientale, Alessandria (2002)
6. Refaeilzadeh, P., Tang, L., Liu, H.: On Comparison of Feature Selection Algorithms. In: Proceedings of AAAI Workshop on Evaluation Methods for Machine Learning II, pp. 34–39. AAAI Press, Vancouver (2007)
7. Schlapbach, A., Bunke, H.: Off-line Handwriting Identification Using HMM Based Recognizers. In: Proc. 17th Int. Conf. on Pattern Recognition, pp. 654–658. IEEE Press, Washington (2004)
8. Bensefia, A., Nosary, A., Paquet, T., Heutte, L.: Writer Identification by Writer's Invariants. In: Eighth Intl. Workshop on Frontiers in Handwriting Recognition, pp. 274–279. IEEE Press, Washington (2002)
9. Shen, C., Ruan, X.-G., Mao, T.-L.: Writer Identification Using Gabor Wavelet. In: Proceedings of the 4th World Congress on Intelligent Control and Automation, vol. 3, pp. 2061–2064. IEEE Press, Washington (2002)
10. Srihari, S.N., Cha, S.-H., Lee, S.: Establishing Handwriting Individuality Using Pattern Recognition Techniques. In: Sixth Intl. Conference on Document Analysis and Recognition, pp. 1195–1204. IEEE Press, Washington (2001)
11. Said, H.E.S., Tan, T.N., Baker, K.D.: Writer Identification Based on Handwriting. Pattern Recognition 33, 149–160 (2000)
12. Bensefia, A., Paquet, T., Heutte, L.: A Writer Identification and Verification System. Pattern Recognition Letters 26, 2080–2092 (2005)
13. Yu, K., Wang, Y., Tan, T.: Writer Identification Using Dynamic Features. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 512–518. Springer, Heidelberg (2004)
14. Tapiador, M., Sigüenza, J.A.: Writer Identification Method Based on Forensic Knowledge. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 555–561. Springer, Heidelberg (2004)
15. He, Z.Y., Tang, Y.Y.: Chinese Handwriting-based Writer Identification by Texture Analysis. In: Proceedings of 2004 Intl. Conference on Machine Learning and Cybernetics, vol. 6, pp. 3488–3491. IEEE Press, Washington (2004)

16. Wirotius, M., Seropian, A., Vincent, N.: Writer Identification From Gray Level Distribution. In: Seventh Intl. Conference on Document Analysis and Recognition, pp. 1168–1172. IEEE Press, Washington (2003)
17. Marti, U.-V., Messerli, R., Bunke, H.: Writer Identification Using Text Line Based Features. In: Sixth Intl. Conference on Document Analysis and Recognition, pp. 101–105. IEEE Press, Washington (2001)
18. Bin, Z., Srihari, S.N.: Analysis of Handwriting Individuality Using Word Features. In: Seventh Intl. Conference on Document Analysis and Recognition, pp. 1142–1146. IEEE Press, Washington (2003)
19. Zois, E.N., Anastassopoulos, V.: Morphological Waveform Coding for Writer Identification. Pattern Recognition 33, 385–398 (2000)
20. Kudo, M., Sklansky, J.: Comparison of Algorithms that Select Features for Pattern Classifiers. Int J. Pattern Recognition 33, 25–41 (2000)
21. Schlapbach, A., Kilchherr, V., Bunke, H.: Improving Writer Identification by Means of Feature Selection and Extraction. In: Eight Intl. Conference on Document Analysis and Recognition, pp. 131–135. IEEE Press, Washington (2005)
22. Pratama, S.F., Muda, A.K., Choo, Y.-H.: Feature Selection Methods for Writer Identification: A Comparative Study. In: Proceedings of 2010 Intl. Conference on Computer and Computational Intelligence, pp. 234–239. IEEE Press, Washington (2010)
23. Srihari, S.N., Huang, C., Srinivasan, H., Shah, V.A.: Biometric and Forensic Aspects of Digital Document Processing. In: Chaudhuri, B.B. (ed.) Digital Document Processing, pp. 379–405. Springer, Heidelberg (2006)
24. Srihari, S.N., Cha, S.-H., Arora, H., Lee, S.: Individuality of Handwriting. J. Forensic Sciences 47, 1–17 (2002)
25. Zhu, Y., Tan, T., Wang, Y.: Biometric Personal Identification Based on Handwriting. In: Intl. Conference on Pattern Recognition, vol. 2, pp. 797–800. IEEE Press, Washington (2000)
26. Cajote, R.D., Guevara, R.C.L.: Global Word Shape Processing Using Polar-radii Graphs for Offline Handwriting Recognition. In: TENCON 2004 IEEE Region 10 Conference, vol. A, pp. 315–318. IEEE Press, Washington (2004)
27. Parisse, C.: Global Word Shape Processing in Off-line Recognition of Handwriting. IEEE Trans. on Pattern Analysis and Machine Intelligence 18, 460–464 (1996)
28. Madvanath, S., Govindaraju, V.: The Role of Holistic Paradigms in Handwritten Word Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 23, 149–164 (2001)
29. Steinherz, T., Rivlin, E., Intrator, N.: Offline Cursive Script Word Recognition – ASurvey. Intl. Journal on Document Analysis and Recognitio 2, 90–110 (1999)
30. Cheikh, F.A.: MUVIS: A System for Content-Based Image Retrieval. PhD Thesis, TampereUniversity of Technology (2004)
31. Vinciarelli, A.: A Survey on Off-line Cursive Word Recognition. Pattern Recognition 35(7), 1433–1446 (2002)
32. Liao, S.X.: Image Analysis by Moment. PhD Thesis, University of Manitoba (1993)
33. Hu, M.K.: Visual Pattern Recognition by Moment Invariants. IRE Transaction on Information Theory 8, 179–187 (1962)
34. Alt, F.L.: Digital Pattern Recognition by Moments. J. the ACM 9, 240–258 (1962)
35. Reiss, T.H.: The Revised Fundamental Theorem of Moment Invariants. IEEE Trans. on Pattern Analysis and Machine Intelligence 13, 830–834 (1991)
36. Belkasim, S.O., Shridhar, M., Ahmadi, M.: Pattern Recognition with Moment Invariants: A Comparative Study and New Results. Pattern Recognition 24, 1117–1138 (1991)
37. Pan, F., Keane, M.: A New Set of Moment Invariants for Handwritten Numeral Recognition. In: IEEE Intl. Conference on Image Processing, vol. 1, pp. 154–158. IEEE Press, Washington (1994)

38. Sivaramakrishna, R., Shashidhar, N.S.: Hu's Moment Invariant: How Invariant Are They under Skew and Perspective Transformations? In: Conference on Communications, Power and Computing, pp. 292–295. IEEE Press, Washington (1997)
39. Palaniappan, R., Raveendran, P., Omatu, S.: New Invariant Moments for Non-Uniformly Scaled Images. Pattern Analysis and Applications 3, 78–87 (2000)
40. Shamsuddin, S.M., Darus, M., Sulaiman, M.N.: Invarianceness of Higher Order Centralised Scaled-Invariants on Unconstrained Handwritten Digits. Intl.J. Inst. Maths. and Comp. Sciences 12, 1–9 (2001)
41. Yinan, S., Weijun, L., Yuechao, W.: United Moment Invariant for Shape Discrimination. In: IEEE Intl. Conference on Robotics, Intelligent Systems and Signal Processing, pp. 88–93. IEEE Press, Washington (2003)
42. Zhang, D.S., Lu, G.: Review of Shape Representation and Description Techniques. Pattern Recognition 37, 1–19 (2004)
43. Chen, C.-C.: Improved Moment Invariants for Shape Discrimination. Pattern Recognition 26, 683–686 (1993)
44. Yu, L., Liu, H.: Efficient Feature Selection via Analysis of Relevance and Redundancy. J. Machine Learning Research, 1205–1224 (2004)
45. Geng, X., Liu, T.-Y., Qin, T., Li, H.: Feature Selection for Ranking. In: 30th Annual Intl. ACM SIGIR Conference, pp. 407–414. ACM Press, Amsterdam (2007)
46. Whitney, A.W.: A Direct Method of Nonparametric Measurement Selection. IEEE Trans. in Computational, 1100–1103 (1971)
47. Pudil, P., Novovicova, J., Kittler, J.: Floating Search Methods in Feature Selection. Pattern Recognition Letter 15, 1119–1125 (1994)
48. Marti, U.-V., Bunke, H.: The IAM Database: An English Sentence Database for Off-line Handwriting Recognition. J. Document Analysis and Recognition 5, 39–46 (2002)
49. Balthrop, J., Forrest, S., Glickman, M.R.: Coverage and Generalization in An Artificial Immune System. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 3–10. Morgan Kaufmann, San Francisco (2002)
50. Palhang, M., Sowmya, A.: Feature Extraction: Issues, New Features, and Symbolic Representation. In: Huijsmans, D.P., Smeulders, A.W.M. (eds.) VISUAL 1999. LNCS, vol. 1614, pp. 418–427. Springer, Heidelberg (1999)
51. Khotanzad, A., Lu, J.H.: Classification of Invariant Image Representations Using a Neural Network. IEEE Trans. on Acoustics, Speech and Signal Processing 38, 1028–1038 (1990)
52. Liu, C.-L., Dai, R.-W., Liu, Y.-J.: Extracting Individual Features from Moments for Chinese Writer Identification. In: Proceedings of the Third Intl. Conference on Document Analysis and Recognition, vol. 1, pp. 438–441. IEEE Press, Washington (1995)
53. Muda, A.K.: Authorship Invarianceness for Writer Identification Using Invariant Discretization and Modified Immune Classifier. PhD Thesis, Universiti Teknologi Malaysia (2009)
54. Liu, H., Setiono, R.: A Probabilistic Approach to Feature Selection - A Filter Solution. In: Intl. Conference of Machine Learning, pp. 319–337. Morgan Kaufmann, Bari (1996)
55. Yu, L., Liu, H.: Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: Proceedings of the Twentieth Intl. Conference on Machine Learning, pp. 856–863. ICM Press, Washington (2003)
56. Rish, I.: An Empirical Study of the Naive Bayes Classifier. In: Intl. Joint Conference on Artificial Intelligence, pp. 41–46. AAAI Press, Vancouver (2001)
57. Breiman, L.: Random Forests. J. Machine Learning, 5–32 (2001)
58. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. J. SIGKDD Explorations 11, 10–18 (2009)