

Invariant behavioural based discrimination for individual representation

Wong Yee Leng¹, Siti Mariyam Shamsuddin², Nor Azman Hashim³

^{1,3}School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Malaysia

²Big Data Centre, Universiti Teknologi Malaysia, Malaysia

Article Info

Article history:

Received Nov 15, 2019

Revised Jun 25, 2020

Accepted Jul 11, 2020

Keywords:

Authorship

Behavioural biometric

Data mining

Discretization

Identification

ABSTRACT

Writer identification based on cursive words is one of the extensive behavioural biometric that has involved many researchers to work in. Recently, its main idea is in forensic investigation and biometric analysis as such the handwriting style can be used as individual behavioural adaptation for authenticating an author. In this study, a novel approach of presenting cursive features of authors is presented. The invariants-based discriminability of the features is proposed by discretizing the moment features of each writer using biometric invariant discretization cutting point (BIDCP). BIDCP is introduced for features perseverance to obtain better individual representations and discriminations. Our experiments have revealed that by using the proposed method, the authorship identification based on cursive words is significantly increased with an average identification rate of 99.80%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Wong Yee Leng,

Applied Computing, School of Computing,

Faculty Engineering, Universiti Teknologi Malaysia (UTM),

81310 Johor Bahru, Johor, Malaysia.

Email: nureilayah@utm.my

1. INTRODUCTION

Pattern identification is commanding countless in the area of engineering like manufacturing, industrial, business including scientific disciplines like artificial intelligence, computer visualisation, remote recognising, ecology, psychology, remedy, and others. One of the well recognized area in pattern recognition is handwriting exploration, which is crucial in biometric and forensic analysis such as writer identification (WI). An author can be identified using individual writing style. Individual writing style has long been considered as individualistic, and author individuality rests on the assumption that each individual has reliable writing style [1-3]. This handwriting must have distinctive feature that could be generalized as individual behavioural features through handwriting shape, and this can be done through identification process. Manual writer identification (WI) based cursive handwriting needs handwriting specialist (graphologist) to discover the individuality writing features of those handwritings accordingly. Normally, features from the original handwritten document will be compared with features from the list of suspects' handwritten documents. These features will be evaluated and compared to obtain features similarity. If these chores are adjusted into computerized system, then the classical procedures of pattern recognition will take place, which include feature extraction and classification.

Many research works have been done to solve identification problem by using image processing and pattern recognition techniques [4-7]. However, from the literature reviews, no research has been done on solving WI problem using Moment Function as features extractors and Discretization methods as

a mechanism to probe the behavioural individuality of each writer based on cursive handwriting. Hence, this research proposes the use of moment function and improved discretization to identify an authorship of authors' cursive handwriting accordingly. A conventional geometric function with united moment invariant is implemented to extract the writers' features. Extensive exploration on the invarianceness of these invariants will be probed to seek the individuality of writing. Subsequently, these invariants features will be discretized to granularly mine these features for identifying the authorship of the writers. Despite the common usage of Discretization in data mining, to the best of our knowledge, no such study has been conducted on Discretization of the invariant behavioural features for cursive handwriting particularly in pattern recognition.

The paper is outlined as follows: The current issues of WI are given in the next section. Following section provides an introduction of moment functions, united moment invariant (UMI) and integrated invariants of aspect invariant scaling (ASI) into UMI for Cursive handwriting. Next section discusses the invarianceness of cursive authorship by the proposed Discretization, followed by the computation, analysis of the proposed method in terms of inter-class and intra-class to illustrate the concept of individuality and discriminability. Following is the section that reveals its implementation and results. Finally, the last section concludes the paper and possible future work.

WI can be counted as a specific kind of vibrant biometric where the characters, shapes and handwriting styles of individual can be used as biometric features for authenticating an identity [8-11]. Typically, WI performed on official papers by a way of signature. However, there is a need to identify a writing style of a documents without signature from a personal such as in threaten letter, writer determination of old or ancient manuscript, and film script (to identify the original idea). The author credentials for questioned handwritten document have a great consequence on the criminal justice system and widely explored in forensic handwriting analysis [1, 3, 12-14]. Despite many researchers in WI, the challenges still arise due to the limitation of human capability in observing and recognizing the style of handwriting. Hence, it has been an inspiration to the researchers to have in depth exploration on this field. The shape or style of cursive writing from one person to another is different and even for one person, it diverse in times. However, everyone has their own style of writing and typically, it is individualistic [1-5]. The feature must be unique, thus can be generalized as person's handwriting regardless of countless writing styles. An individual's writing style has its own particular texture and structure [12]. Each handwriting shape is slightly dissimilar for same author and relatively different for dissimilar authors. This is known as intra writer class for the same writer and inter writer class for different writers. These extracted features are entailed to be classified for group or class identification. In the concept of Pattern Recognition, it is widely depends on feature extraction, classification and learning schemes as described in [13, 14]. Those techniques are important and are required in order to obtain true authorship of handwritten. In between, the process of eliminating, extracting and choosing the exact features of a person's handwritten are not an easy task in the area of pattern recognition prior to classification, where those best extracted features will be grouped into specific categories. However, it is an open question whether the extracted features are optimal or near-optimal to identify the author. Features mining may include irrelevant features, and useless for classification and sometimes degrading the performance of a classifier [15]. The features may not be independent of each other or even redundant. Moreover, there may be some features that do not provide any useful information for the task of WI. Hence, mining significant features are very important in order to identify the writer, moreover to improve the identification rate. Therefore, the objective of this paper is to explore indiscriminate distinctive behaviour features of written cursive words style by implementing integrated moment functions to acquire the features from handwriting, and discretized these data in order to represent them significantly. The basic ideas about moment functions as feature extraction in our study will be well illustrated in the next section.

2. RESEARCH METHOD

2.1. United moment invariant (UMI) and aspect scaling invariant (ASI) for cursive extraction

Moment function has been used in diverse fields ranging from mechanics and statistics to pattern recognition and image understanding. The use of moments in image analysis and pattern recognition was inspired by Hu [16] and Alt [17]. Hu first presented a set of seven-tuplet moments that invariant to position, size, and orientation of the image shape. A good shape descriptor should be able to find perceptually similar shape where it usually means rotated, translated, scaled and affined transformed shapes. Furthermore, it can tolerate with human in comparing the image shapes. Therefore, Yinan [18] derived united moment invariants (UMI) that can be applied in all conditions with a good set of discriminate shapes features. It effectively discriminates the shape of image on both region and boundary in discrete and continuous condition. UMI was derived based on the geometric moment invariant (GMI) and the improve moment invariant (IMI) [19].

GMI is usable for region representation in discrete condition but high in computational times for boundary representation. Thus, Yinan *et al.*, [18] proposed the above UMI that has been proven as a good technique for feature extraction task. Unfortunately, UMI technique uses scaling factor by Hu which was already proven to have some drawbacks in terms of scaling factor. Therefore, the alternative scaling factor of aspect invariant moment (Aspect) by Feng and Keane [20] is used in this study. It obtained better invariant features without size normalisation. A fusion formulation of the scaling factor of Aspect [20] into the UMI [18] algorithm is applied in this study to extract the global word shape features from both region and in boundary representation, in discrete and continuous condition for better individual features. UMI are best method use to discriminate handwriting features and applicable in any discrete condition as described by Hu [16], which considers normalized central moments as shown below (refer to [16] for detail formulations of GMI):

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^2}, \quad p+q=2,3,\dots \quad (1)$$

$$\mu_{pq} = \int \int_{-\infty-\infty}^{\infty \infty} (x-x_0)^p (y-y_0)^q f(x,y) dx dy, \quad p,q=0,1,2,3\dots$$

with (2) in discrete form. Central and normalized central moments are summarized as below:

$$\mu'_{pq} = \rho^{p+q} \mu_{pq},$$

$$\eta'_{pq} = \rho^{p+q} \eta_{pq} = \frac{\rho^{p+q}}{\mu_{00}^2} \mu_{pq} \quad (2)$$

where ρ is a scaling factor. The enhanced moment invariant technique by Chen [21] is specified as follows:

$$\eta'_{pq} = \frac{\mu_{pq}}{(\mu_{00})^{p+q+1}} \quad (3)$$

Equation (1) to (3) consists of features μ_{pq} . By disregarding the features of μ_{00} and ρ , UMI can be presented as below:

$$\begin{aligned} \theta_1 &= \frac{\sqrt{\phi_2}}{\phi_1} & \theta_2 &= \frac{\phi_6}{\phi_1 \phi_4} & \theta_3 &= \frac{\sqrt{\phi_5}}{\phi_4} & \theta_4 &= \frac{\phi_5}{\phi_3 \phi_4} \\ \theta_5 &= \frac{\phi_1 \phi_6}{\phi_2 \phi_3} & \theta_6 &= \frac{(\phi_1 + \sqrt{\phi_2}) \phi_3}{\phi_6} & \theta_7 &= \frac{\phi_1 \phi_5}{\phi_3 \phi_6} & \theta_8 &= \frac{(\phi_3 + \phi_4)}{\sqrt{\phi_5}} \end{aligned} \quad (4)$$

where ϕ_i are Hu's moment invariants and each constituent of ϕ_i involves μ_{pq} , (see Yinan [18] and Hu[16]). On the other hand, based on Feng [20], GMI introduced by Hu [16] have numerous disadvantages and only invariant with an equal scaling image. Therefore, Feng [20] proposed aspect invariant moment (AIM) for imageries of inadequate scaling size by integrating the ideas of moment invariants that are efficient in solving different scaling in both directions of x and y. The invariant scaling that was proposed by Feng [20] is called aspect scaling invariant (ASI), and is given as:

$$\eta_{pq} = \frac{\mu_{00}^{\frac{p+q+2}{2}}}{\mu_{20}^{\frac{p+1}{2}} \mu_{02}^{\frac{q+1}{2}}} \mu_{pq} \quad (5)$$

However, in this study, we used the integration of ASI and UMI (AUMI) to extract the features as given below. The detail of the integration is described in Muda [22].

$$\theta_1 = \frac{\sqrt{\phi_2}}{\phi_1} = \frac{\sqrt{(\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2}}{\eta_{20} + \mu_{02}} = \frac{\sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{\mu_{20} + \mu_{02}} \quad (6)$$

2.2. The proposed biometric invariant discretization cutting point (BIDCP) for authorship invarianness

As mentioned previous, feature extraction and learning scheme play significant role in determining and identifying the performance of handwritten authorship. Many approaches have been conducted in extracting and selecting the meaningful features. However, the issues of identifying the behavioural structures that are optimum or less-optimum are still infancy. In this study, the concept of Discretization is proposed to granularly mining the extracted invariants' features for better individual representation in cursive writer identification. It is used as an important role in leading to better identification for WI. Based on previous studies, classification approaches that work the best for pre-processing process are the one that integrated with discretization [23]. It discretized globally all the features of the writers. In other words, the continuous extracted features are discretized to attain the uniqueness of authors' individuality for better data representation [24]. Hence, the proposed discetization so-called biometric invariant discretization cutting point (BIDCP) is applied to the class information that is assigned to each writer to assure the distinctiveness and individual personality perseverance. Interval and representation features are formed based on each writer. If the features of two different writers are quite closed to each other or the values are the same, then comparable intervals for these two groups are generated. The novel approach here transforms those feature vectors into better behavioural representation without changing any characteristics. BIDCP first compute the feasible intervals for the given datasets. The minimum (v_{min}) and the maximum (v_{max}) of the features vectors (fv) for a writer are obtained. A cutting point of feature vectors that starts from the minimum (v_{min}) and ends with the maximum (v_{max}). In this study, The Interval is used to define the cutting points for the representation value of each writer. For this, we denote that the interval as the width of the bin as calculated in (7).

$$Interval = \frac{(v_{max} - v_{min})}{n} \quad (7)$$

The entire bins is created equivalent to the overall feature vectors that represent one word image. Therefore, the entire invariant vectors in moment invariant function are preserved to its actual features. Each bin is then approximated with upper and lower values as demonstrated in (8) and (9). Each feature vector that falls in the interval of upper to lower approximation range is defined with a single representation value (rv), as illustrated in (10) which is the improvised version from previous Azah's Discretization [22]. Instead of taking the range between the interval, the (rv) is considered by taking the midpoint of the AV_{upper} and AV_{lower} .

$$AV_{upper} = v_{min} + Interval \quad , \quad (8)$$

$$AV_{lower} = AV_{upper} \quad , \quad (9)$$

$$rv = (AV_{upper} + AV_{lower}) / 2 \quad (10)$$

In this study, there are nine features that represent one word of a writer. Thus, each writer is corresponded by nine bins or intervals. For the interval one to eighth, the representation value (rv) of a writer is represented as the features vectors in the range of ($AV_{lower} \leq fv < AV_{upper}$). Here, AV_{upper} is not included in the first until eighth bin of a writer because it is used as base value to construct new approximation value for next bin. This range specifies boundary to each word written by the same writer. If there are two different words written by same writer that have close or same invariant features that fall within this range, hence there will be the same representation value for these two words of the same writer. This is because the values are calculated based on each writer. Therefore, the intention of the proposed algorithm is not to change the usual characteristic of writer but just to symbolize the original invariant behavioural feature into better feature representation. However, for the last interval, it is defined by the representation value (rv) of features vectors in the range of ($AV_{lower} \leq fv \leq AV_{upper}$) where the equality sign include the upper approximation as well. This range represents the writer's of the exact

class. If the features fall within this range, they are symbolized as the writer’s from the same class. Otherwise, it is considered as features from other class. Overall, feature values that fall within this both ranges are known as discretized features. With this improved procedure in computing intervals as illustrated in (10), the estimated representation feature values are more close to the actual biometric behavioural features distribution (true feature values). This preserves the discriminative power of the original features and enhances the statistical distinctiveness between individuals. Figure 1 illustrates the discretization process for writer 1. Each word image is represented as vector of nine discretized biometric invariant behavioural features.

From Figure 1, it states that each individual has its own unique representation features, which denote the main characteristic of each writer accordingly. This delineates the concept of individuality and discriminability assurance where each person has its own handwriting style. To further validate the effectiveness of the proposed Discretization, the individual obtained features are tested with author invariance analysis to evaluate the concept of individuality in handwriting.

Bin 0	Lower	Upper	Representation Value
1	2.8	4.02222	3.41111
2	4.02222	5.24444	4.63333
3	5.24444	6.46667	5.85556
4	6.46667	7.68889	7.07778
5	7.68889	8.91111	8.3
6	8.91111	10.1333	9.52222
7	10.1333	11.3556	10.7444
8	11.3556	12.5778	11.9667
	12.5778	13.8	13.1889

(a)

<i>f</i> 1	<i>f</i> 2	<i>f</i> 3	<i>f</i> 4	<i>f</i> 5	<i>f</i> 6	<i>f</i> 7	<i>f</i> 8	<i>f</i> 9	Class
13.1889	13.1889	13.1889	13.1889	13.1889	12.2000	13.1889	13.1889	3.41111	W1
3.41111	13.1889	13.1889	12.2010	13.1889	13.1889	13.1889	13.1889	13.1889	W1
13.1889	13.1889	13.1889	13.1889	10.7444	11.9667	11.9667	11.9667	10.7444	W1
13.1889	13.1889	11.9667	11.9667	11.9667	10.7444	13.1889	13.1889	13.1889	W1
12.2111	12.2111	10.9899	12.2111	10.9899	12.2111	12.2111	12.2111	12.2111	W2
12.2111	12.2111	12.2111	11.2389	12.2111	12.2111	12.2111	12.2111	12.2111	W2
11.2389	11.3024	12.2111	10.9899	12.2111	11.2389	12.2111	13.1141	12.2111	W2
12.2111	12.2111	12.2111	13.1141	14.2000	11.2389	11.2389	12.2111	13.1141	W2
.....									
.....									

Discriminatory values of individual 1 for signature is 13.1889

Discriminatory values of individual 2 for signature is 12.2111

(b)

Figure 1. (a) BIDCP process for W1, (b) Examples discretized features for cursive words words performed by different writers; W1 and W2

3. RESULTS AND ANALYSIS

3.1. Analysis of cursive handwriting invarianceness

As mentioned previous, feature extraction and learning schemes play significant role in determining the invarianceness of individual in the perspective of moment functions. They can be signified as images perseverance irrespective to its transformations. Mathematically, Tomas Suk and Flusser [25] describe invariant I as a functional features on the space of all permissible image functions which will not modify its value that below deficiency operator D , which fulfils the condition of $I(f) = I(D(f))$ for related image function f , which known as invariance. Another appropriate operator I , as significant as invariance, is discriminability. For substances belong to another classes, I need to have significantly different values. Therefore, in our study, we define authorship invarianceness in WI as low similarity deviation for same writer (called as intra writer class) and high similarity deviation for different writers (called as inter writer class) depending on wrting shape. This is due to the distinctiveness of each person writing style which is called as authorship invarianceness. The essential process of individual identification in WI is to look for comparable characteristic of wrting style based on the nearest unidentified individual wrting style in the record. This be able to solve by applying handwriting distinctiveness, and this can be achieved by conducting the computation of intra writer class and inter writer class. The objective of intra writer class and

inter writer class is to find the nearest characteristic by obtaining the lowest mean absolute error (MAE) for authorship invarianceness. In the context of WI, Intra Writer Class (same writer) should provide lowest Mean Absolute Error compared to Inter Writer Class (different writers), irrespective to writing styles and word shapes. The mean absolute error (MAE) function is showned as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |(x_i - r_i)| \quad (11)$$

Note that the similarity deviation for Inter Writer Class (different writers) should be higher than Intra Writer Class (same writer) in the concept of authorship invarianceness. These similarity deviations represent the data by discerning the individual features into category. The fundamental idea is to obtain objects that can be easily categorized into one of the interval; hence this is labelled as discretisation approach. These similarity errors can be allied into discretisation to exemplify the data by discriminating the individual features into appropriate class. To the best of our knowledge, no study has been conducted on implementing discretization process in Writer Identification for cursive handwriting. Therefore, we propose this approach in our research study for authorship invarianceness. Tables 1 to 4 show the similarity result using MAE for intra-class (same writer) and inter-class (different writers) on Chinese characters and signature. Tables 1 and 2 illustrate the authorship for Intra Writer Class (same writer) on Chinese character, which is smaller compared to Inter Writer Class (different writers) for the similar word. Same results goes for inter writer class on dissimilar words like 天, 成, 天 and 中 where its similarities result is greater than intra writer class in authorship invarianceness. Interestingly, same results is found in longer characters like signatures.

Table 1. MAE for intra writer (same) and inter writer (different) on same Chinese characters

Chinese Character	Intra Writer (1 writer)	Inter Writer (15 writers)	Inter Writer (30 writers)	Inter Writer (45 writers)
天	0.50552	1.02844	1.1837	1.19672
成	0.48471	0.87482	0.72221	0.71014
中	0.50578	0.90938	1.01647	1.27493
国	0.39898	0.6738	0.53648	0.51642

Table 2. MAE for intra writer (same) and inter writers (different) on various of Chinese characters

Various Chinese Characters	Intra Writer (1 writer)	Inter Writer (15 writers)	Inter Writer (30 writers)	Inter Writer (45 writers)
15 characters	0.66492	1.17773	1.07744	0.99323
30 characters	0.53283	0.95377	0.99738	0.83671
45 characters	0.36721	0.75129	0.69983	0.67382



Tables 3 and 4 show the result of MAE after the proposed Discretization on same and different signature characters respectively. Again, it is proven that the writer writing style for signature, where MAE value for intra writer class is lower compared to inter write r class, regardless of simple or complex word like  or . This is due to the competence of Discretisation in mining the features class without any complications in terms of dimension and difficulties of the handwriting shape. Thus, this authorship invarianceness analysis confirms this novel approach is able to extract the unique behaviour features for individual identification.

Table 3. MAE for intra writer (same) and inter writers (different) on same signature words

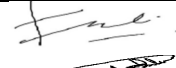


Signature	Intra Writer (1 writer)	Inter Writer (15 writers)	Inter Writer (30 writers)	Inter Writer (45 writers)
	0.94738	1.2658	1.19991	1.04991
	1.07401	1.37381	1.23233	1.18471
	0.75113	1.48601	1.49847	1.49389

Table 4. MAE for intra writer (same) and inter writer (different) on various of signature words

Various Signatures	Intra Writer (1 writer)	Inter Writer (15 writers)	Inter Writer (30 writers)	Inter Writer (45 writers)
15 signatures	0.87611	1.58612	1.57822	1.6383
30 signatures	0.93684	1.47828	1.37769	1.13293
45 signatures	1.03843	1.63522	1.57362	1.4079

3.2. Experimental results

This section investigates the improvement of identification performance based on handwriting using the proposed discretization by utilizing the dissimilar types of discretization methods, determined on a variety of classification methods. The comparison results are examined by using Chinese characters and signature biometric modalities from In-house multimodal biometric database. The experiment is tested on hundred subjects, where each subject contributes four samples of different style of Chinese handwritten characters and signatures. Numerous type of words are extracted, using integrated moment functions to signify the word in terms of feature vector. Since we are using moment functions images are not necessarily to be converted to binary representation. These feature vectors have gone through the discretization process prior to classification. Since the major contribution of our study is focus on our proposed discretization, hence, our comparisons are bounded to existing built-in discretization methods in rough set tool for data analysis (ROSETTA).

We are concentrating more on the effectiveness of the discretization mechanism in mining the granularity of the extracted features using moment functions. The experiments are conducted to assess the identification performance by performing different discretization methods like CAIM, CACC, ChiM, Chi2, ExtChi2, and Khiops discretization as well as classification methods of ROSETTA. These include Johnson algorithm, Holte IR Algorithm, genetic algorithm and exhaustive algorithm. Meanwhile, Biometric invariant discretization cutting point (BIDCP) is the proposed Discretization algorithm. 100 writers with 800 various handwritten Chinese character and signature images are extracted for each moment function as adopted in this study. The number of data for each word is different for each writer in our house biometric database. Therefore, different numbers of data can be prepared for each type of word in feature extraction task for each writer. About 7,200 invariant feature vectors of each technique are divided into training and testing data set in the identification task. The results of the experiments for data set using six discretization and four classification methods are summarized and reported into a single Table 5.

Table 5. Performance of four classification methods on various types of discretization algorithms

Classification	K-NN		C45		NB		SVM	
	Train	Test	Train	Test	Train	Test	Train	Test
CAIM	55.001	78.893	78.991	74.009	70.888	70.892	60.000	70.999
CACC	60.895	69.733	72.893	66.897	65.092	60.001	65.983	69.923
ChiM	65.342	68.324	70.372	70.324	76.001	75.833	70.324	70.999
Chi2	79.558	81.444	80.004	79.526	80.532	70.860	77.666	70.432
ExtChi2	78.433	77.922	78.334	78.339	81.000	80.999	81.788	79.123
Khiops	59.328	80.788	79.052	78.832	86.732	79.782	78.003	78.000
BIDCP (Proposed)	99.213	99.700	99.800	98.920	99.555	98.999	99.592	98.859

Table 5 shows the four types of classification methods with application of the proposed BIDCP discretization that gives the best response to the available biometric modalities. As it can be seen here, the implementation of proposed discretization on data set yields a higher average accuracy rate (over 98.5%) than other six discretization algorithms namely CAIM, CACC, ChiM, Chi2, ExtChi2, and Khiops respectively. The combination of proposed discretization with K-NN, C-45, NB and SVM classification on training datasets successfully achieved the best performance with the average accuracy rate of 99.213%, 99.500%, 98.555%, and 99.192% respectively. Whereas, for testing dataset, the performance of K-NN, C-45, NB and SVM classification also yields a higher performance with the average accuracy of 99.700%, 98.920%, 98.999%, and 98.859% after applying the proposed approach on the data sets. Meanwhile, the second best on the combination of CAIM and four classification methods on biometric datasets, while the worst for the CACC method. It clearly shows that our BIDCP discretized features give higher identification rates for all samples.

From these experiments, we found that the identification rates using discretized features are significantly greater compared to non-discretized features (original features). This is due to the features invarianceness and features discriminability that has been improved using our proposed discretization

algorithm as well as other discretized methods. The features are assembled explicitly in same class (interval) and corresponding to the same individual with similar representation value. This representation value portrays the uniqueness of each writer respectively. This leads to lower variation for features in intra-class concept, and higher variation in inter-class concept. Hence, Authorship invarianceness and Author discriminability have been presented accordingly with better handwriting individuality.

4. CONCLUSION AND FUTURE WORK

In this studies, we proposed a novel method of presenting features discriminability by implementing discretization process prior to classification phase. The main goal of proposing discretization approach in the area of pattern recognition framework is to represent features in a granular form to obtain better individual representation. Our discretized data shows the characteristics of individuality in handwriting are well represented. The similarities of the same writers are also minimized between features, thus, leads to better identification accuracy. We have presented the findings of our generalized features for handwriting individuality using moment function. In future work, we will further mining these generalized features for better identification in forensic document analysis.

ACKNOWLEDGEMENTS

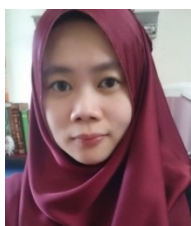
This research is sponsored partly by the School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia. Authors would especially like to thank UTM Big Data Centre and Soft Computing Research Group (SCRG), UTM for their excellent cooperation and contributions to improve this paper.

REFERENCES

- [1] S. N. Srihari, C. Huang, H. Srinivasan, V. A. Shah, "Biometric and forensic aspects of digital document processing," Digital Document Processing, B. B. Chaudhuri (ed.), Springer, 2006.
- [2] L. Xiaohong, L. Yuanyuan, "Handwriting identification: Challenges and solutions," *J Forensic Sci Med*, vol. 4, pp. 167-73, 2018.
- [3] B. Ameur and T. Hatem, "Validity of Handwriting in Biometric Systems," *PRAI 2018: Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence*, pp. 5-10, 2018.
- [4] A. Rehman, S. Naz, M. I. Razzak, "Writer identification using machine learning approaches: A comprehensive review," *Multimedia Tools Appl.*, pp. 1-43, Sep. 2018.
- [5] C. Adak, B. B. Chaudhuri and M. Blumenstein, "An Empirical Study on Writer Identification and Verification From Intra-Variation Individual Handwriting," in *IEEE Access*, vol. 7, pp. 24738-24758, 2019.
- [6] P. Pandey, K. R. Seeja, A. Somani, S. Srivastava, A. Mundra, S. Rawat, "Forensic Writer Identification with Projection Profile Representation of Graphemes," *Proceedings of First International Conference on Smart System Innovations and Computing. Smart Innovation Systems and Technologies*, vol. 79, pp. 129-136, 2018.
- [7] Rokiah Rozita Ahmad, Maslina Darus, Siti Mariyam Shamsuddin and Azuraliza Abu Bakar, "Pendiskretan Data Set Kasar Menggunakan Ta'akulan Boolean (Rough Set Data Discretization using Boolean Reasoning)," *Jurnal Teknologi Maklumat & Multimedia*, vol. 1, pp. 15-26, 2004.
- [8] M. Bulacu, L. Schomaker, "Combining multiple features for text-independent writer identification and verification," in *10th International Workshop on Frontiers in Handwriting Recognition (IWFHR 2006)*, 23-26 October, La Baule, France, pp. 281-286, 2006.
- [9] Gazzah S., E. Ben Amara N., "Writer Identification Using Modular MLP Classifier and Genetic Algorithm for Optimal Features Selection," In: Wang J., Yi Z., Zurada J.M., Lu B.L., Yin H. (eds) *Advances in Neural Networks - ISNN 2006*. ISNN 2006. Lecture Notes in Computer Science, vol. 3972. Springer, Berlin, Heidelberg, 2006.
- [10] P. Pandey, K. R. Seeja, "Forensic writer identification with projection profile representation of graphemes," *Proc. 1st Int. Conf. Smart Syst. Innov. Comput.*, pp. 129-136, 2018.
- [11] F. Cloppet, V. Eglin, V. C. Kieu, D. Stutzmann, N. Vincent, "ICFHR2016 competition on the classification of medieval handwritings in latin script," *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 590-595, Oct. 2016.
- [12] A. Durou, I. Aref, S. Al-Maadeed, A. Bouridane, E. Benkhelifa, "Writer identification approach based on bag of words with OBI features," *Inf. Process. Manage.*, vol. 56, no. 2, pp. 354-366, 2019.
- [13] A. A. Ahmed, H. R. Hasan, F. A. Hameed, O. I. Al-Sanjary, "Writer identification on multi-script handwritten using optimum features," *Kurdistan J. Appl. Res.*, vol. 2, no. 3, pp. 178-185, 2017.
- [14] S. Fiel, F. Kleber, M. Diem, V. Christlein, G. Louloudis, N. Stamatopou-Los, B. Gatos, "ICDAR 2017 competition on historical document writer identification (Historical-WI)," *2017 14th International Conference on Document Analysis and Recognition*, Kyoto, pp. 1377-1382, Nov. 2017.
- [15] J. J. Miller, R. B. Patterson, D. T. Gantz, C. P. Saunders, M. A. Walch, J. Buscaglia, "A set of handwriting features for use in automated writer identification," *J. Forensic Sci.*, vol. 62, no. 3, pp. 722-734, 2017.
- [16] M. K. Hu, "Visual pattern recognition by moment invariants," *IRE Transaction on Information Theory*, vol. 8, no. 2, pp. 179-187, Feb. 1962.

- [17] F. L. Alt, "Digital pattern recognition by moments," *Journal of the ACM (JACM)*, vol. 9, no. 2, pp. 240-258, 1962.
- [18] T. H. Reiss, "The revised fundamental theorem of moment invariants, Pattern Analysis and Machine Intelligence," *IEEE Transactions*, vol. 13, no. 8, pp. 830-834, Aug. 1991.
- [19] S. O. Belkasim, M. Shridhar, M. Ahmadi, "Pattern recognition with moment invariants: a comparative study and new results," *Pattern Recognition*, vol. 24, no. 12, pp. 1117-1138, 1991.
- [20] P. Feng, and M. Keane, "A new set of moment invariants for handwritten numeral recognition, Image Processing," in: *ICIP-94, IEEE International Conference*, vol. 1, pp. 154-158, Nov. 1994.
- [21] C.-C. Chen, "Improved moment invariants for shape discrimination," *Pattern Recognition*, vol. 26, no. 5, pp. 683-686, May 1993.
- [22] A. K. Muda, S. M. Shamsuddin, and M. Darus, "Embedded scale united moment invariant for identification of handwriting individuality," in: *ICCSA 2007 International Conference on Computational Science and Its Applications, Computational Science and Its Applications – ICCSA 2007*, Springer Verlag, pp. 385-396, 2007.
- [23] U. Stańczyk, "On Unsupervised and Supervised Discretisation in Mining Stylometric Features," In: Gruca A., Czachórski T., Deorowicz S., Harężlak K., Piotrowska A. (eds) *Man-Machine Interactions 6. ICMMI 2019 International Conference on Man-Machine Interactions. Advances in Intelligent Systems and Computing*, Springer, Cham, vol. 1061, 2020.
- [24] A. K. Muda, S. M. Shamsuddin and M. Darus, "Invariants Discretization for Individuality Representation in Handwritten Authorship," *International Workshop on Computational Forensic (IWCF 2008), LNCS 5158*, Springer Verlag, pp. 218- 228, 2008.
- [25] T. Suk, J. Flusser, "Affine moment invariants Generated by graph method," *Pattern Recognition*, vol. 44, no. 9, pp. 2047-2056, Sep. 2011.

BIOGRAPHIES OF AUTHORS



Wong Yee Leng obtained her Master degree and Doctor of Philosophy from Universiti Teknologi Malaysia (UTM) in department of Computer Science. Previously, she was a research assistant at the Department of Software Engineering, UTM, concerned with the Research and Development in the areas of Speech Recognition and Multimodal Biometric System. Currently, she is a lecturer at Applied Computing, Faculty Engineering at UTM and a member of Soft Computing Research Group. Her research interests include Big Data Computing, Artificial Intelligence, Pattern Recognition, Multimodal Biometric and Machine Learning.



Siti Mariyam Shamsuddin obtained her Master degree in Mathematics from Fairleigh Dickinson University (FDU) New Jersey USA and Doctor of Philosophy in Artificial Intelligence from Universiti Putra Malaysia (UPM). Previously, she was a Head of R & D Cluster of Engineering and ICT, Head Department of Computer Graphics and Multimedia from 2001 until 2006, and Head, Soft Computing Research Group. Currently, she is a Director, UTM Big Data Centre, Skudai Johor Bahru which focuses on Data Science and Big Data Analytics ranging from R & D products and services. Her research interests include Big Data Computing, Machine Learning, Soft Computing, Pattern Recognition, Biometrics Systems and Intelligent Graphics.



Nor Azman Ismail obtained his Master Degree in Information Technology from Universiti Kebangsaan Malaysia and Doctor of Philosophy from Loughborough University, United Kingdom. He was a Deputy Director of Corporate Affairs (Web Director) since April 2009. Currently, he is Associate Professor in School of Computing, Faculty Engineering at Universiti Teknologi Malaysia and holding a post as Associate Chair (Research & Academic Staff). His research interests include Human Computer Interaction (HCI), Image Retrieval and Webometric in improving the state of research, practice, and education.