

RESEARCH

Open Access

Improvement in twins handwriting identification with invariants discretization

BO Mohammed* and SM Shamsuddin

Abstract

One of the most popular areas of study in pattern recognition which has now become the centre of many researchers' attention is Writer Identification. A more recent development in the area is Twins Handwriting Identification which has now become not only an important, but also widely popular area of study especially in the fields of forensic research and biometrical identification. In terms of biometrical identification, it is known that a pair of twins may share various similar traits genetically. Forensic evidence can be easily obtained from handwriting samples. Therefore, in order to achieve reliable and accurate identification based on handwriting, it is important for the similarities in the writing traits of a pair of twins to be differentiated. In identifying an individual, handwriting style can be analyzed to allow the implicit representation of the unique hidden features of the individual's handwriting. Said unique features can help in identifying the writer of the text which can be essential when identifying the writer between a pair of twins. Previous studies in authorship identification were highly concentrated in the study of the classification task as well as features extraction. However, the issue of the similarities in the traits of a pair of twins' handwriting were not taken into account thus, leaving a high possibility of degrading the performance of the classification process. Therefore, in order to achieve better input for the classification task, this article will discuss an additional process which can better represent an individual's personal features through the transformation of the similarities via discretization protocol. The additional process can help improve the level of identification for Individuality of Handwriting of a pair of twins.

Keywords: writer identification, unique representation, authorship invarianceness, twins handwriting, discretization

1 Introduction

Despite the advancement and technological achievement of the current age, documents are still printed on paper and widely exchanged, hence the need for Writer Identification (WI). WI helps to properly identify the writer of a handwritten document. Shapes and styles of an individual's handwriting contain hidden personal traits of the writer which can contribute to the process of handwriting identification in dynamic biometric study. The biometric features are used to identify the identity of the writer [1-4]. This is also applicable in the case of identifying the writer between a pair of twins. Through studying the signature, WI is commonly used to authenticate legal paper. WI can also be used in identifying the authorship of documents without signature, such as letters of threat, historical, or ancient manuscripts and

other documents only containing handwritten text without the person's signature. The current technology allows WI to be performed even with the use of limited samples of handwriting. In the field of handwriting analysis for forensic purposes, WI holds great importance and is widely used on evidence to be used in the courtroom [5-8]. Therefore, the many issues and challenges in Twins Handwriting Identification need to be given attention for further investigation.

The Twins Handwriting Identification is a quite popular area of research in pattern recognition and computer vision fields as it, in some situations, provides the only means of discovering the real writer of a written text out of a group of people [9,10].

Proven through previous studies on twins' biometric identification which include the studies on the discriminability between the fingerprints of a pair of twins [11], DNA analysis [12], computational discriminability analysis on the fingerprints of a pair of twins [13],

* Correspondence: bayancomputer@yahoo.com
Soft Computing Research Group, Faculty of Computer Science and Information Systems, University Technology Malaysia, Johor, Malaysia

coefficient values shown in individual sets which form of unique code for an individual's face [14], natural physiological traits is unchanging throughout an individual's life. However, unlike an individual's psychological traits which remain constant, the association between an individual's handwriting with the individual's behavioral nature allows the handwriting of the individual to change according to the changes in their behavior and provide a strong reasoning behind the study of handwritings [15].

Distinguishing the handwriting of a pair of twins is a challenge in the area of biometric study. Throughout the years, it has been noticed that the unique features of an individual is embedded in the individual's handwriting. Therefore, through studies and with the current status of knowledge, various techniques have been developed and further improved to properly study handwriting samples [15].

Through studying a pair of twins' handwriting, the writer can efficiently be distinguished. This form of study has been proven to be more complex compared to studying the handwriting of non-twins. This is due to the fact that the resemblance of a pair of twins' characteristics is also shown in their writing manners which generate similar features in their handwriting. There are two stages of the identification phase: the analysis of the individual feature as well as the identification of the features with similarities and the capture of the features. The results of the stages will be the functions and are computerized with the help of the classical method of identification.

In identifying the author of a handwritten text, previous studies on WI have shown more interest in the tasks of feature extraction classification. However, most work did not focus on the additional step which in this article will be focusing. The additional step aims to provide better representation for the input which will be used in the classification process. Better representation of the input can help in a way that the classification task can be done more quickly and accurately for the real writer to be more accurately identified especially in the case of handwriting identification of a pair of twins. The features extracted in the feature extraction process show that the handwriting of a pair of twins has very similar representations which causes a problem once the input is used in the classification process as similarities will lower the accuracy of the classification task. This article will provide the discussion on the additional step of transformation where the closely similar representation of features are transformed into clearer and better representations which can represent each twin.

2 Individuality of Twins Handwriting

An individual's nature can be seen through his or her handwriting and the hypothesis as mentioned in [16-18]

stated that a person's individuality in writing shows through the fact that said person has a consistent form of handwriting. These figures are samples of handwritten texts from three pairs of twins. Figure 1 shows a sample of handwritten text of the same characters and Figure 2 samples with different characters. It can be seen that the shape of the writings are only slightly different when the author is the same writer in a pair of twins while have more defined difference for different authors in a pair of twins although the height of the writings are similar. This difference is 'Individuality of Handwriting' in which the difference in handwriting is still evident even between a pair of twins. This form of individuality is measurable by the variances where the feature of the writer (intra-class) has to be of lower value than that of different writers (inter-class) [19-21]. Individual features are considered good and acceptable if the features have the lowest similarity error for one author in a pair of twins (intra-class) and highest similarity error for both authors in a pair of twins (inter-class) [19]. Therefore, individual features need to be acquired from the samples of handwritten texts to be able to identify the authorship of the text when the identification involves a pair of twins. This concept of handwriting individuality was defined and discussed in [22] as authorship invarianceness.

3 Unique representation

Features are used as input in the identification process used by the classifier; therefore, it is important to obtain good and reliable features in order to achieve accurate and clear identification. It is common for the features extracted from the features extraction process to be used directly by the classifier in the classification process. However, in the case of handwriting identification involving a pair of twins, it is not suitable to be used directly as the representation of the individual features of a pair of twins are usually very closely similar which causes the intra-class variance to be large and the inter-class variance to be small. Hence, in order to improve the invarianceness of authorship, another process can be added before the features are used in the classification process. This study implements the Invariant Discretization Technique from [19] on samples of twins' handwriting. The technique is meant to reduce the intra-class variance of the features while increasing the inter-class variance. Figure 3 shows an overview of the study which led to the need of this additional procedure to be performed for the identification of a pair of twins' handwriting to be improved.

4 Feature extraction

Macro-features which represent the global characteristics of the writing habit and style of an individual can

Twins number 1		Twins number 2		Twins number 3	
Twins 1 a	Twins 1 b	Twins 2 a	Twins 2 b	Twins 3 a	Twins 3 b
a	a	a	a	a	a
a	a	a	a	a	a
a	a	a	a	a	a
a	a	a	a	a	a

Figure 1 Same character between twins.

be captured and extracted from an entire document [10,15]. These macro-features are used in this study for the purpose of identifying the writer between a pair of twins. Thirteen macro-features including the 11 initial features stated in [10,17] are used in this study. The 11 features include the entropy of grey values, the binarization threshold, number of black pixels, number of interior contours, number of exterior contours, contour slope components consisting of number of horizontal, number of positive, number of vertical and number of negative, the average height as well as the average slant. Only eight features are used in the experiments of this study which are the entropy of grey values, the binarization threshold, number of black pixels, number of interior contours, and number of exterior contours, average height, average slant, and average stroke width. Macro-features have been chosen for the experiments because of the global characteristics captures by the features which can present the writer's individuality in terms of writing style and habit [18]. Detailed descriptions of the macro-feature algorithm are provided in [15-17].

5 Discretization

In classification, the problem in focus is usually the training instances. The set of instances which have distinct, descriptive features are usually categorized into

classes. In the discretization process, the transformation of the continuous features forms discrete partitions with a certain number of intervals. A lower and an upper boundary represent the range of each interval. As there are many ways in representing the continuous features, certain important points are needed. The first point is to determine the number of intervals for each discrete partition. The number is usually selected at random. Second, the boundaries are decided for the intervals. There are several known methods for discretization including Equal Information Gain, Maximum Entropy, and Equal Interval Width. Another method proposed in [19], the Invariants Discretization method, has however been proven more efficient in providing higher accuracy and success rates of identification. The Invariants Discretization method is a supervised method. The method starts by searching the appropriate intervals to represent the writer's information. The upper and lower boundaries are then set for each interval. The number of intervals for an image must be the same as the number of the feature vectors.

The individual's uniqueness can be computed according to each writer and the preservation of the information help ease the task of classification. This discretization process proved to be beneficial in terms of nonlinear representation [23] and through the set of

Twins number 1		Twins number 2		Twins number 3	
Twins 1 a	Twins 1 b	Twins 2 a	Twins 2 b	Twins 3 a	Twins 3 b
a	a	a	a	a	a
b	b	b	b	b	b
d	d	d	d	d	d
e	e	e	e	e	e
h	h	h	h	h	h

Figure 2 Different character between twins.

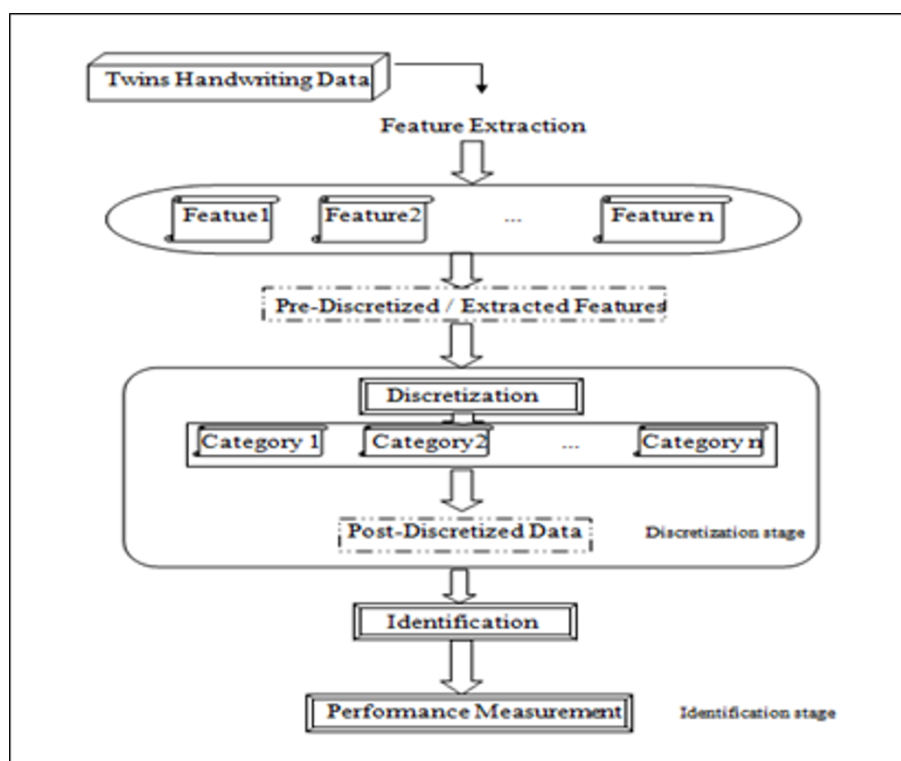


Figure 3 Framework of discretization for WI for a pair of twins.

intervals, interpretation can easily be done by humans [24]. Reducing the amount of data also helps the computation process to be done quicker [25,26]. According to the authors of [19], use of post-discretized data provided higher level of identification compared to using pre-discretized data. The result of the study showed that through the application of the discretization method on the proposed integrated Moment Invariant, higher accuracy can be achieved.

5.1 Discretization protocol

An appropriate number of intervals with a representation value representing the extracted feature are calculated in the discretization process. The representation value, called discretized feature vector, is where the 'generalized unique feature' for each individual feature is obtained. The generalized feature illustrates the hidden features of an individual's writing style. Then minimum and maximum range of the data for each writer is divided into intervals which can be called 'cuts' of equal sizes in order to obtain an interval. The number of feature vector columns of the extracted features defines the number of intervals.

The example shows eight feature vector columns obtained from the macro-feature technique. Each interval is given a lower and an upper approximation and

one representation value represents each interval. In the supervised discretization, the value is calculated based on the writer class. An invariant feature vector which falls into an interval will have the interval's representation. Therefore, writers with closely similar invariant feature vectors will have similar intervals for the two classes. The information and characteristics of a writer are not affected by the Discretization algorithm. The algorithm only presents the invariant feature vector originally extracted from the feature extraction process in a standard representation with generalized features. Figure 4 shows an illustration of the discretization algorithm.

Invariant discretization requires the writer class information for the discretization process. The calculation of the range of intervals in the invariant discretization line uses the minimum (v_{\min}) invariant feature vector and the maximum (v_{\max}) invariant feature vector (if v) of the writer. A line for a writer starts with the minimum (v_{\min}) invariant feature vector and ends with the maximum (v_{\max}) invariant feature vector. The interval is the average of the invariant discretization line when divided by the number of invariant feature vector column. The calculation of the interval's width (wd) is as follows:

$$wd = (v_{\max} - v_{\min}) / f \quad (1)$$

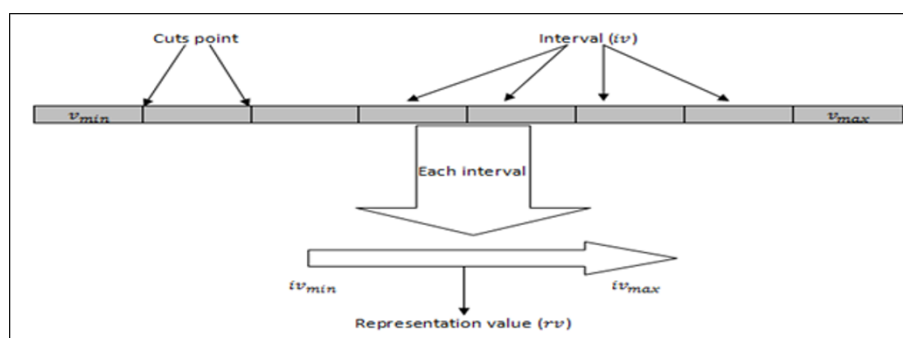


Figure 4 Invariant discretization line [19].

where v_{\min} is the minimum value of invariant feature vector for a writer; v_{\max} is the maximum value of invariant feature vector for a writer; and f is the number of invariant feature vector column.

The interval in an invariant discretization line has cut points which are defined by the width. The invariant feature vector in an interval will have the interval's representation value. The representation value (rv) of each interval is the average of interval. It is calculated as $rv = (iv_{\max} - iv_{\min})/2$. Intervals 1 to 7 are represented with the representation value of the invariant feature vector within if $v \geq iv_{\min}$ and if $v < iv_{\max}$ while the invariant feature vector within if $v \geq iv_{\min}$ and if $v \leq iv_{\max}$ is put under the category of the last interval. The representation value, known as discretized feature vector, is a representation of the unique features in an individual's writing. Figures 5 and 6 show the transformation of the invariant feature vector into discretized feature vector for pre- and post-discretized data, respectively. It can be seen that the discretization algorithm provides discretized feature vector that shown clear illustration of an individual's unique features, even between a pair of twins.

6 Simulation result

Two experiments were conducted in this study: the experiment on the authorship invarianceness for the handwriting sample of a pair of twins and the evaluation of the accuracy of identification between the pair of twins. The first experiment was conducted in order to prove that the discretization technique improves the variance of the intra-class (same writer in a pair of twins) and inter-class (both writers in a pair of twins) features. The second experiment was conducted to evaluate the discretization in terms of improving the performance of the identification of the writer between a pair of twins using the Rosetta Toolkit [27] and artificial neural network (ANN). The data used for the experiments were from the collection of 390 data samples

obtained from 13 pairs of identical twins from the Sulaimania University, Iraq.

6.1. Authorship invarianceness between twins

Through the use of the Mean Absolute Error (MAE) function, the authorship invarianceness can be measured. Figure 7 presents an example of the MAE calculation. For each twin, there are 15 images of handwriting samples. Features 1 to 8 are the features extracted to represent a character. The character's invarianceness and the reference image (the first image) are given by the MAE value [19]. Small errors indicate that the image is close or similar to the reference image. The average MAE is calculated from the overall result.

$$MAE = \frac{1}{n} \sum_{i=1}^f |x_i - r_i| \quad (2)$$

where n is the number of images; x_i is the current image; r_i is the reference image or location measure; f is the number of features; i is the feature column of image.

The calculation for the authorship invarianceness for post- and pre-discretized feature vectors can be achieved through analyzing the intra-class and the inter-class of the MAE value. The result of the analysis show that the use of post-discretized feature vector feature provides improved authorship invarianceness compared to the use of pre-discretized feature vector as the intra-class MAE value using the post-discretized feature vector is smaller and the inter-class MAE value is higher than that of the pre-discretized feature vector. Low MAE value for intra-class indicates that the features for a single writer in a pair of twin are similar while the high value of MAE for inter-class indicates that the features of the handwriting of each twin is different from the another. The hypothesis is therefore proven correct and the discretization process is deemed able to improve the authorship invarianceness with the standard representation of the individual's unique features presented clearly to help identify the writer between a pair of twins.

5.153	1.72	2.382	0.709	1.8	2.656	8.576	2.5	15a
3.511	3.07	3.308	0.668	1.8	5.890	5.255	2.2	15b
3.957	1.94	2.506	0.682	0.9	3.359	8.978	2.1	15b
4.810	1.63	2.162	0.702	1.4	2.875	3.905	2.2	15a
6.702	1.96	2.636	0.684	4.4	6.015	1.351	2.6	15a
3.408	2.57	2.958	0.694	0.9	5.046	0.204	2.3	15a
6.520	1.65	1.937	0.707	0.9	2.234	5.760	2.8	15a
4.934	1.78	2.665	0.692	1.8	2.796	1.467	2.4	15b
7.593	1.29	2.194	0.698	1.8	0.125	8.296	2.7	15b
5.827	2.39	2.864	0.694	1.8	5.046	2.413	2.3	15b
6.021	1.59	2.460	0.703	2.3	0.203	2.510	2.4	17a
5.029	1.85	2.566	0.694	1.8	3.078	5.014	2.5	17a
2.297	1.77	2.553	0.703	1.4	1.328	9.648	2.3	17a
5.395	1.70	2.535	0.703	1.4	2.031	7.697	2.6	17a
6.841	2.33	3.095	0.688	2.7	3.359	9.920	2.7	17a
7.645	2.41	2.945	0.700	1.8	5.750	4.322	2.5	17a
4.975	2.10	2.931	0.703	1.4	2.734	7.487	2.6	17b
6.024	2.02	2.902	0.698	2.2	2.156	4.012	2.7	17b
2.404	2.58	3.087	0.694	1.9	5.625	0.202	2.3	17b
2.856	2.21	2.875	0.682	2.7	4.625	1.928	2.3	17b
5.680	3.28	3.786	0.680	2.3	7.234	6.840	2.5	17b
5.315	2.37	2.818	0.721	5.4	7.015	1.157	2.6	17b
6.947	3.27	3.349	0.711	5.3	0.515	6.973	2.8	17b
5.136	2.44	3.293	0.705	0.9	6.734	4.568	2.5	18a
4.208	3.22	3.316	0.702	1.8	9.546	3.104	2.1	18a
2.934	2.86	2.725	0.717	3.6	8.703	4.467	1.9	18a
6.168	2.71	3.253	0.717	1.8	7.718	8.584	2.7	18a
3.922	2.45	3.255	0.700	1.8	6.875	4.461	2.4	18a
8.571	2.40	2.610	0.721	1.9	5.765	4.285	2.8	18a
6.281	2.80	3.094	0.723	4.2	7.390	3.140	2.3	18a
8.689	4.83	3.880	0.682	6.8	4.406	5.844	2.5	18b
3.654	2.40	3.103	0.678	3.2	4.984	1.827	2.3	18b
6.029	2.34	3.058	0.654	1.8	4.062	0.014	2.6	18b
5.261	3.09	3.298	0.676	3.6	5.328	7.130	2.4	18b
6.029	2.92	3.357	0.660	1.8	5.328	9.014	2.7	18b
6.292	3.27	3.517	0.654	3.2	7.093	6.646	2.5	18b
6.049	3.26	3.520	0.666	1.8	5.750	6.024	2.6	18b

Figure 5 Example of pre-discretized twins datasets.

Figures 8 and 9 show the comparison of the authorship invarianceness for the macro-feature technique with post- and pre-discretized data, respectively.

Figures 8 and 9 show the results which describe an individual's unique features where even between a pair of twins, the uniqueness is evident. As the value of the MAE for intra-class (same writer in a pair of twins) is lower than the value of the MAE for inter-class (both writer in a pair of twins), it satisfies the concept that states that there are traits of individualities in the handwriting of a pair of twins. Using post-discretized feature vector, the individual features can be better illustrated compared to using pre-discretized feature vector. The post-discretized data should have a lower MAE value than the pre-discretized data for intra-class (same writer in the same twins), and the post-discretized data should give a higher MAE value when compared to the pre-discretized data for inter-class (both writer in the same twins). Furthermore, the results shown in Figure 10 also show that the use of post-discretized data improved the MAE value for inter-class.

6.2. Identification performance evaluation rough set classifier

In the classification task, whether it is to lessen the computation time, or to minimize classification errors, any method may be chosen based on its efficiency and ability to complete the task as required. In this article, rough set theory was chosen for its ability to deal with the upper and lower approximation concepts of the set which provides a way of classifying objects in noisy or incomplete condition.

In [28], it is stated that the boundary region of a set is represented by the set difference between its upper and lower approximations. Figure 11 illustrates the concept of rough set theory. Figure 12 on the other hand shows the approximation role of Rough set concept.

With the use of the Rosetta (Rough Set Toolkit) as suggested in [27], an experiment was conducted to evaluate the performance of the writer identification between a pair of twins which uses both the post- and pre-discretization techniques. The experiment takes into account the additional step used in the study for the purpose of Twins' Handwriting Identification which

5.7023	2.0254	2.0254	0.7998	2.0254	0.7998	2.0254	2.0254	17a
4.4767	2.0254	2.0254	0.7998	2.0254	3.2511	4.4767	2.0254	17a
2.0254	2.0254	2.0254	0.7998	0.7998	0.7998	9.3792	2.0254	17a
5.7023	2.0254	2.0254	0.7998	0.7998	2.0254	8.1536	2.0254	17a
6.9279	2.0254	3.2511	0.7998	3.2511	3.2511	9.3792	3.2511	17a
8.1536	2.0254	3.2511	0.7998	2.0254	5.7023	4.4767	2.0254	17a
4.4767	3.2511	3.2511	0.7998	3.2511	6.9279	9.3792	2.0254	17a
4.4767	2.0254	3.2511	0.7998	2.0254	6.9279	6.9279	2.0254	17a
4.4667	2.0297	3.2482	0.8112	0.8112	3.2482	6.9037	2.0297	17b
5.6852	2.0297	3.2482	0.8112	2.0297	2.0297	4.4667	3.2482	17b
2.0297	2.0297	3.2482	0.8112	2.0297	5.6852	0.8112	2.0297	17b
3.2482	2.0297	3.2482	0.8112	3.2482	4.4667	2.0297	2.0297	17b
5.6852	3.2482	3.2482	0.8112	2.0297	6.9037	6.9037	2.0297	17b
5.6852	2.0297	3.2482	0.8112	5.6852	6.9037	0.8112	2.0297	17b
6.9037	3.2482	3.2482	0.8112	5.6852	0.8112	6.9037	3.2482	17b
5.3963	3.0251	3.0251	0.6538	0.6538	6.5819	4.2107	3.0251	18a
8.9532	1.8394	3.0251	0.6538	1.8394	5.3963	4.2107	3.0251	18a
4.2107	3.0251	3.0251	0.6538	1.8394	7.7676	6.5819	1.8394	18a
6.5819	3.0251	3.0251	0.6538	1.8394	7.7676	8.9532	3.0251	18a
4.2107	3.0251	3.0251	0.6538	1.8394	8.9532	3.0251	1.8394	18a
3.0251	3.0251	3.0251	0.6538	3.0251	8.9532	4.2107	1.8394	18a
4.2107	3.0251	3.0251	0.6538	1.8394	6.5819	4.2107	1.8394	18a
6.5819	3.0251	3.0251	0.6538	4.2107	7.7676	3.0251	1.8394	18a
5.5428	3.0856	3.0856	0.6283	1.8569	5.5428	5.5428	3.0856	18b
9.2287	4.3142	4.3142	0.6283	6.7714	4.3142	5.5428	3.0856	18b
3.0856	1.8569	3.0856	0.6283	3.0856	5.5428	1.8569	1.8569	18b
5.5428	1.8569	3.0856	0.6283	1.8569	4.3142	0.6283	3.0856	18b
5.5428	3.0856	3.0856	0.6283	3.0856	5.5428	6.7714	1.8569	18b
5.5428	3.0856	3.0856	0.6283	1.8569	5.5428	9.2287	3.0856	18b
6.7714	3.0856	3.0856	0.6283	3.0856	6.7714	6.7714	3.0856	18b
6.7714	3.0856	3.0856	0.6283	3.0856	4.3142	3.0856	3.0856	18b

Figure 6 Example of post-discretized twins datasets.

helped in increasing the accuracy of the identification through the process of discretization. A total of 390 data samples, divided into 2 datasets; namely, training and testing data, were used for the classification task.

In order to achieve a more reliable and accurate performance with the use of the discretization method, the Cross Validations (CV) in [27] were implemented on the post- and the pre-discretized data. The number of

the folds was specified by the number of the CV iterations. The experiments in this study were done with 10, seven, and fivefold CV iterations. The process of discretization was done based on the Invariant Discretization method by Azah Kamilah.

Two experiments with 70% training data, 30% testing data and 60% training data, 40% testing data were completed (10, 7, 5 cross validation). The CV process









Image	Feature 1	Feature 2	Feature 3	Feature4	Feature 8	MAE
	4.9750	2.1000	2.9310	0.7030	2.6000	-
	6.0240	2.0200	2.9020	0.6980	2.7000	0.4077
	2.4040	2.5800	3.0870	0.6940	2.3000	0.9461
	2.8560	2.2100	2.8750	0.6820	2.3000	0.7571
	5.6800	3.2800	3.7860	0.6800	2.5000	0.5940
	5.3150	2.3700	2.8180	0.7210	2.6000	1.0235
.....
	3.6760	2.7100	3.0700	0.7050	2.5000	0.3694
	4.4020	3.0500	3.5650	0.6840	3.5000	0.8287
Average of MAE							0.6197

Figure 7 Example of MAE calculation.

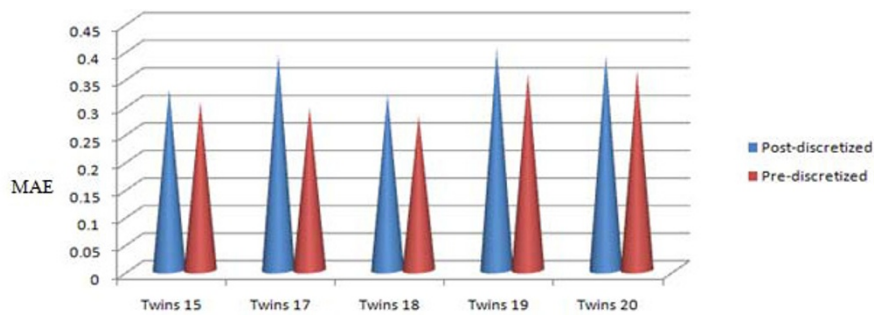


Figure 8 Authorship invarianceness comparison for intra-class (same writer in a pair of twins).

provides the experimental results as shown in Tables 1 and 2. The results are then evaluated and are as visualized in Figures 13 and 14.

Through the results shown in both Tables 1 and 2, it can be concluded that the use of post-discretized data can result in higher accuracy when compared to the use of pre-discretized data. Thus, the use of post-discretized data can significantly improve the performance of Twins' Handwriting Identification.

6.3. Identification performance evaluation with artificial neural network classifier

The ANN classifier is used on both types of the Twins datasets in order to achieve the main goal of the research. In this article, ANN is used to classify the between- and within-writer distances while minimizing misclassification errors. ANNs have several desirable properties: sound statistical procedure, practical software implementation of the Bayesian (optimal) procedure, no presumptions about the nature of the data (unlike other classifiers), and they let us tap into the full multivariate nature of the data and enable us to use a nonlinear discrimination criterion. In this research, we used a 3-layered network: an input layer with eight units and a hidden layer with five units. Figure 15 shows the ANN architecture of this research.

Three experiments were conducted with a varied number of training data and testing data where the first experiment used 70% training data and 30% testing data from a combination of pre-discretized and post-discretized datasets. The second experiment was conducted with the use of 60% training data and 40% testing data. ANN was used for the training process. With the use of the classification matrix, the overall accuracy of identification was calculated from each training and testing dataset.

The results of both the experiment using 70 and 60% training data are as summarized in Table 3. Through the results, it can be noted that the use of post-discretized data can provide an overall identification rate with the Average Accuracy (%) of above 90.0%. The use of pre-discretized data on the other hand has lower identification rate which is below 60.0%. This proves that better identification and higher level of accuracy can be achieved with the use of post-discretized datasets.

7 Conclusion

It can be suggested through the results showing the value of MAE in Section 6.1 that the invarianceness of the authorship between a pair of twins was improved with the use of post-discretized feature vector for both intra-class (same writer in a pair of twins) and inter-

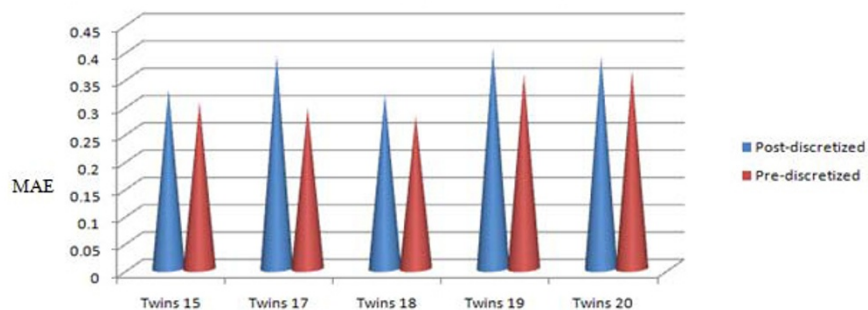


Figure 9 Authorship invarianceness comparison for inter-class (both writer in a pair of twins).

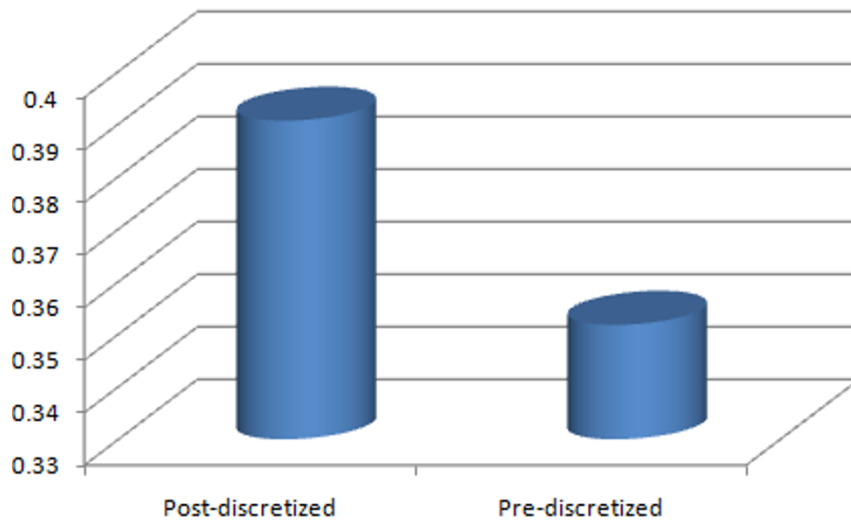


Figure 10 Authorship invarianceness comparison for inter-class for 390 image Handwritings between 13 pairs.

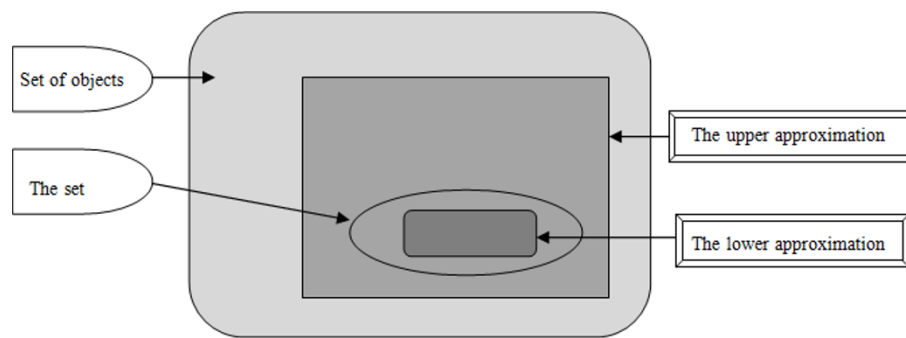


Figure 11 Rough set theory [28].

class (each writer in a pair of twins) when compared to the use of pre-discretized feature vector. This satisfies the concept of Individuality of Handwriting even in terms of Twins' Handwriting Identification where the

concept requires that the intra-class MAE value must be smaller than the inter-class MAE value regardless of the character used for the experiment. The discretization process provided post-discretized feature vector which

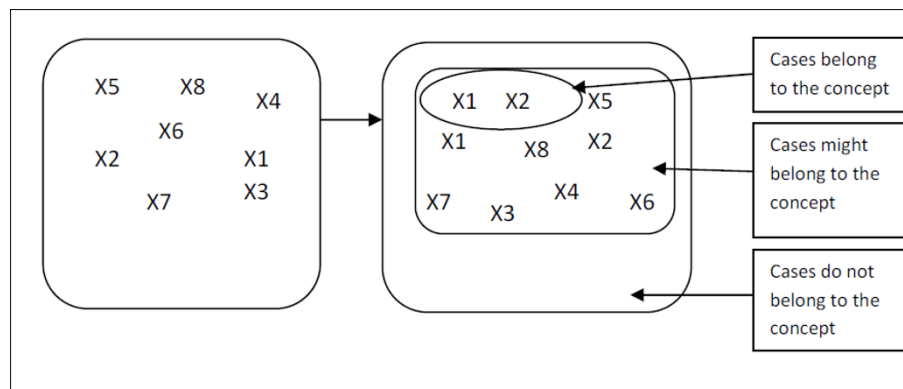


Figure 12 Approximation role in rough set theory [28].

Table 1 Identification rates using different CV with 70% training data and 30% testing data

ROSETA built-in methods an reductions	Genetic algorithm		Holte 1R algorithm		Exhaustive algorithm	
Datasets	Pre-Dis	Post-Dis	Pre-Dis	Post-Dis	Pre-Dis	Post-Dis
Tenfold CV						
Mean (%)	10.32	100.0	06.70	100.0	10.32	100.0
Median (%)	11.53	100.0	05.12	100.0	11.53	100.0
StdDev (%)	04.64	00.00	04.20	00.00	04.64	00.00
Maximum (%)	15.38	100.0	17.94	100.0	15.38	100.0
Minimum (%)	02.56	100.0	02.56	100.0	02.56	100.0
Sevenfold CV						
Mean (%)	09.31	100.0	04.65	100.0	09.31	100.0
Median (%)	10.90	100.0	03.63	100.0	10.90	100.0
StdDev (%)	03.57	00.00	01.44	00.00	03.57	00.00
Maximum (%)	14.54	100.0	07.27	100.0	14.54	100.0
Minimum (%)	03.63	100.0	03.50	100.0	03.63	100.0
Fivefold CV						
Mean (%)	10.87	100.0	06.48	100.0	10.87	100.0
Median (%)	11.68	100.0	07.79	100.0	11.68	100.0
StdDev (%)	03.18	00.00	03.91	00.00	03.18	00.00
Maximum (%)	14.28	100.0	11.68	100.0	14.28	100.0
Minimum (%)	06.32	100.0	02.53	100.0	06.32	100.0

can properly represent and illustrate the individuality of each writer. It proves that the concept of Individuality of Handwriting in WI where each writer has his or her own style of writing with differs even between a pair of twins. The standard representation of the features of each individual consists of small intra-class variance and

large inter-class variance when compared to the invariant feature vectors originally extracted through the features extraction process. As proven in Sections 6.2 and 6.3, this contributes to the higher accuracy of identification for each individual's handwriting. Therefore, it can be concluded that through the analysis of authorship

Table 2 Identification rates using different CV with 60% training data and 40% testing data

ROSETA built-in methods an reductions	Genetic algorithm		Holte 1R algorithm		Exhaustive algorithm	
Datasets	Pre-Dis	Post-Dis	Pre-Dis	Post-Dis	Pre-Dis	Post-Dis
Tenfold CV						
Mean (%)	10.32	100.0	06.70	100.0	10.32	100.0
Median (%)	11.53	100.0	05.12	100.0	11.53	100.0
StdDev (%)	04.64	00.00	04.20	00.00	04.64	00.00
Maximum (%)	15.38	100.0	17.94	100.0	15.38	100.0
Minimum (%)	02.56	100.0	02.56	100.0	02.56	100.0
Sevenfold CV						
Mean (%)	09.31	100.0	04.65	100.0	09.31	100.0
Median (%)	10.90	100.0	03.63	100.0	10.90	100.0
StdDev (%)	03.57	00.00	01.44	00.00	03.57	00.00
Maximum (%)	14.54	100.0	07.27	100.0	14.54	100.0
Minimum (%)	03.63	100.0	03.50	100.0	03.63	100.0
Fivefold CV						
Mean (%)	10.87	100.0	06.48	100.0	10.87	100.0
Median (%)	11.68	100.0	07.79	100.0	11.68	100.0
StdDev (%)	03.18	00.00	03.91	00.00	03.18	00.00
Maximum (%)	14.28	100.0	11.68	100.0	14.28	100.0
Minimum (%)	06.32	100.0	02.53	100.0	06.32	100.0

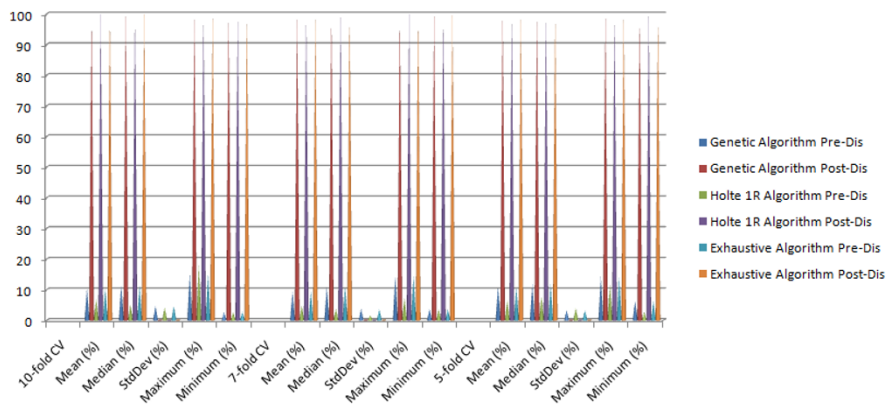


Figure 13 Visualization of divergence level between pre-discretized and post-discretized twins datasets using 70% training and 30% testing data.

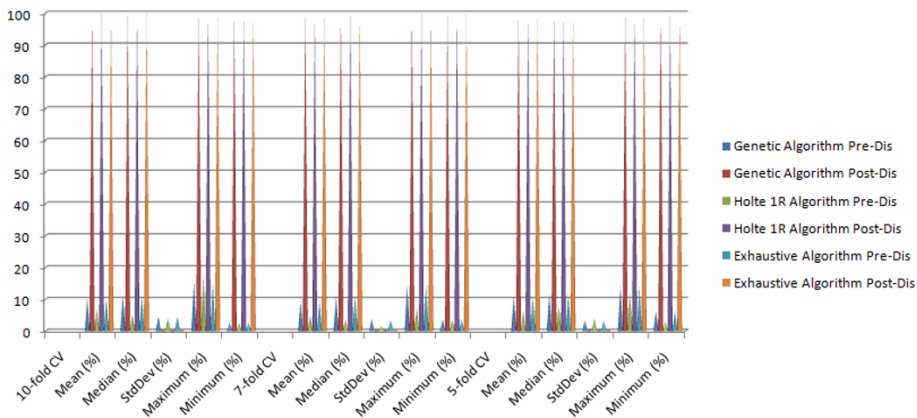


Figure 14 Visualization of divergence level between pre-discretized and post-discretized twins datasets using 60% training and 40% testing data.

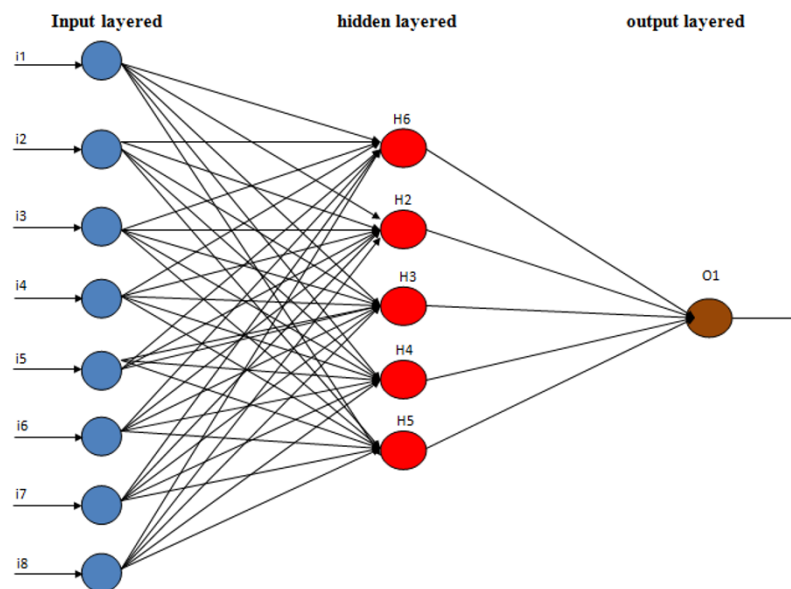


Figure 15 Architecture of ANN.

Table 3 Comparisons of identification rates with different training and testing datasets with ANN

Technique	Accuracy (%), 50% Training Data 50% Testing Data	Accuracy (%) 60% Training Data 40% Testing Data	Accuracy (%) 70% Training Data 30% Testing Data	Datasets
ANN	58.2418	58.9744	59.2308	Pre-Dis
	95.6044	90.3846	93.8462	Post-Dis

invarianceness, the application of the discretization technique should be further explored in the domain of Twins' Handwriting Identification.

Competing interests

The authors declare that they have no competing interests.

Received: 27 December 2011 Accepted: 28 February 2012

Published: 28 February 2012

References

1. SN Srihari, C Huang, H Srinivasan, VA Shah, in *Biometric and Forensic Aspects of Digital Document Processing*, ed. Chaudhuri BB, Digital Document Processing Springer, 379–405 (2006)
2. M Tapiador, JA Siguenza, Writer identification method based on forensic knowledge. in *First International Conference on Biometric Authentication, ICBA, 2004* 555–561 (2004)
3. Y Kun, W Yunhong, T Tieniu, Writer identification using dynamic features, in *Biometric Authentication: First International Conference, ICBA 2004*, Hong Kong, China, 512–518 (15–17 July 2004)
4. Z Yong, T Tieniu, W Yunhong, Biometric personal identification based on handwriting. *Proc 15th International Conference on Pattern recognition, Barcelona, Spain.* **2**, 797–800 (2000)
5. M Somaya, M Erman, K Dori, M Fatma, Writer identification using edge-based directional probability distribution features for arabic words. in *IEEE/ACS International Conference on Computer Systems and Applications, AICCSA Doha* 582–590 (2008)
6. R Niels, L Vuurpijl, L Schomaker, Automatic allograph matching in forensic writer identification. *Int J Pattern Recogn Artif Intell IJPRAL.* **2**(1), 61–81 (2007)
7. V Pervouchine, G Leedham, K Melikhov, Handwritten character skeletonisation for forensic document analysis, in *Proceedings of the 2005 ACM Symposium on Applied Computing*, Santa Fe, New Mexico, USA, 754–758 (2005)
8. K Franke, M Koppen, A computer-based system to support forensic studies on handwritten documents. *Int J Doc Anal Recogn.* **3**(4), 218–231 (2001). doi:10.1007/PL00013565
9. R Plamondon, G Lorette, Automatic signature verification and writer identification state of art. *Pattern Recogn.* **22**, 107–131 (1989). doi:10.1016/0031-3203(89)90059-9
10. SN Srihari, Computational methods for handwritten questioned document examination. Ph.D U.S Department of Justice (2010)
11. AK Jain, S Prabhakar, S Pankanti, On the similarity of identical twin finger prints. *Pattern Recogn.* **35**(1), 2653–2663 (2002). doi:10.1016/S0031-3203(01)00218-7
12. RJ Rubucki, BJ McCue, KJ Duffy, KL Shepard, SJ Shepherd, JL Wisecarver, Natural DNA mixtures generated in fraternal twins in Utero. *J Forensic Sci.* **46**(1), 120–125 (2001)
13. Y Liu, NS Sargur, *A Computational Discriminability Analysis on Twins Fingerprints*. Springer, Heidelberg 43–54 (2009)
14. M Rycchilk, W Stankiewicz, M Moezynski, Method of numerical analysis of similarity and differences of face shape of twins. in *Proceeding of ICBME.* **23**, 1854–1857 http://www.springerlink.com (2009)
15. S Sargur, H Chen, S Harish, S Vivek, On the discriminability of the handwriting of twins. *J Forensic Sci.* **53**(2), 430–446 (2008). doi:10.1111/j.1556-4029.2008.00682.x
16. SN Srihari, S-H Cha, H Arora, S Lee, Individuality of handwriting: a validation study. in *Sixth IAPR International Conference on Document Analysis and Recognition, Seattle* 106–109 (2001)
17. SN Srihari, S-H Cha, H Arora, S Lee, Individuality of handwriting. *J Forensic Sci.* **47**(4), 1–17 (2002)
18. SN Srihari, C Huang, H Srinivasan, VA Shah, *Biometric and Forensic Aspects of Digital Document Processing*. Digital Document Processing Springer, London 379–405 (2006)
19. AK Muda, SM Shamsuddin, A Ajith, Improvement of Authorship invarianceness for individuality representation in writer identification. *Neural Netw World.* **3**(10), 371–387 (2010)
20. G Leedham, S Chachra, Writer identification using innovative binarised features of handwritten numerals. in *Proceeding of Seventh International Conference of Document Analysis and Recognition.* **1**, 413–416 (2003)
21. EN Zois, V Anastassopoulos, Morphological waveform coding for writer identification. *Pattern Recogn.* **33**(3), 385–398 (2000). doi:10.1016/S0031-3203(99)00063-1
22. AK Muda, SM Shamsuddin, M Darus, Invariants discretization for individuality representation in handwritten authorship. in *International Workshop on Computational Forensic (IWCF 2008), LNCS 5158, Springer* 218–228 (2008)
23. G Agre, S Peev, On supervised and unsupervised discretization. *CIT: Cybern Inf Technol.* **2**(2), 43–57 (2002)
24. H Liu, F Hussain, CL Tan, M Dash, Discretization: an enabling technique. *Data Min Knowl Disc.* **6**, 393–423 (2002). doi:10.1023/A:1016304305535
25. P Prachya, R Thanawin, W Kitsana, DCR: discretization using class information to reduce number of intervals (2009). QIMIE/PAKDD 17–28 (2009)
26. GJ Hwang, F Li, in *A Dynamic Method for Discretization of Continuous Attributes, IDEAL, LNCS 2412* Springer, Berlin 506–511 (2002)
27. A Øhrn, J Komorowski, ROSETTA: a rough set toolkit for analysis of data. in *Third International Joint Conference on Information Sciences, Durham, NC.* **3**, 403–407 (1997)
28. Z Pawlak, Rough sets. *Int J Comput Inf Sci.* **11**, 341–356 (1982). doi:10.1007/BF01001956

doi:10.1186/1687-6180-2012-48

Cite this article as: Mohammed and Shamsuddin: Improvement in twins handwriting identification with invariants discretization. *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:48.

Submit your manuscript to a SpringerOpen journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com