Discretization Method to Generalize Features for Authors' Recognition

Intan Ermahani A. Jalil[1], Siti Mariyam Shamsuddin[2], Azah Kamilah Muda[3], Sabrina Ahmad[4], Mohd Sanusi Azmi[5]

[1,3,4,5]Computational Intelligence and Technologies Lab, Centre for Advanced Computing Technology, Fakulti Teknologi Maklumat Dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia
[2]Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia
ermahani@utem.edu.my

ABSTRACT
Features reconstruction or representation to improve the performance accuracy rate is a crucial procedure for handwriting image analysis in the area of authors' recognition. The process of feature extraction usually led to features redundancy with high similarity between features or classes. This will cause the problem of lower performance from classifier models that will have difficulties to differentiate the authors. This study proposes to reconstruct the formulation of discretization method of Equal Width Binning (EWB) to simplify the process. Discretization method is proposed to represent the features and infused the generalization factor into features that are being generated. This is aimed to improve the performance accuracy for authors' recognition. This study deploys the Higher-Order United Moment Invariant (HUMI) and Edge based Directional (ED) feature extraction method to generate Global and Local features respectively from the handwriting images. These generated features were reconstructed by discretization method by representing them with each unique general features. The proposed discretization method has succeeded to improve the average performance of Global Discretized Features that performs at 99.81% and Local Discretized Features achieved up until 99.89%. This shows that the discretization method has managed to improve the features' performance in determining the authors' characteristic.

Keywords: *Accuracy, classification, discretization, recognition.*

I. INTRODUCTION

Data transformation and representation for the purpose of improving predictive accuracy are usually done to compute better input values for any classification model in any field of study when dealing with large amount of data [1]–[4]. Discretization gives the advantages of reducing and simplifying complex data by deploying the generalization factor to the datasets to improve the performance accuracy rate [5]–[6]. This is aimed to transform the representation of each data into invariant discretization that is easier to use and learn by classifiers. There are currently several studies proposing that this method of discretization either supervised or unsupervised has improved their work with increased classification accuracy [7]–[12]. They have managed to present their data in a simple, consistent and accurate way. The study by [10]–[11] has proposed the method of Equal Width Binning (EWB) as a supervised discretization method for the independent offline handwriting dataset. This method is also been proposed by [6] to be included in the new scheme of writer identification that proposed the features ranking procedure being done after discretization procedure towards the handwriting image for writer identification.

Besides, [13] has implemented the work proposed by [10]–[11] into the research of twins' handwriting to determine the variation of style between the twins to identify their individual unique features. Their work has produced the result as high as 97.66% of accuracy rate on average for discretized features as compared to the non-discretized features. In other, the adaptation of this Equal Width Binning (EWB) method into the handwriting of Chinese characters is presented by [9] for the purpose of supervised discretization and contributing towards the use of discretized local features for writer identification.

The research that has discussed on which are the best discretization methods and their effect on classification is done through systematic literature with extensive experiments and analysis by [14]. On the other hand, [15] has long ago discussed the comparison between several supervised and unsupervised discretization methods for continuous features and have reached some conclusion that the unsupervised discretization method has significantly improved the induction algorithms for Naïve Bayes and also C4.5. This shows that the discretization method is influential towards contributing to the increase in performance of some features and contributing to the improvement factor for induction algorithms of some classifiers.

This research proposes to implement the discretization method on the handwriting image for author's recognition by reconstructing the Equal Width Binning (EWB) method into another simple formulation to represent the discretization method. This paper is aligned as the following: Section II covers the reconstruction of the proposed method in Methodology while Section III describes in detail the process to generate the invariant discretization based on the proposed method. Thus, Section IV discusses about the results and lastly Section V is the conclusion for this study.

II. METHODOLOGY

The variation of feature vectors extracted by any feature extraction methods may produce some similarity in feature values. Feature extraction procedure may produce massive amount of data [18]–[20]. This may effects the performance of the classifiers and resulting to lower performance accuracy as the classifiers' model could not differentiate the characteristics of the author's handwriting image from the input feature vectors.

Thus, this study proposed the reconstruction formulation of discretization method by [10]–[11] to simplify the procedure in order to infuse the generalization factor for all feature vectors where each features are represented by their own unique values. The task of discretization is aimed to produce the range of data to become the unique universal representation of each pre-discretized data values.

Discretization process starts with finding the minimum and maximum values from the entire features for each class. Both values are required to determine the range of features data whereby it presents the width of each interval or bin based on the number of features data to be computed in the range of values produced by the class.

As supposed the number of features data is represented by k, minimum value by min, maximum value by max and the width of each interval or bin is provided by w.

$$w = (max - min)/k \qquad (1)$$

Thus, the production of each interval or bin is as follows:

$$min + w, min + 2w, \ldots, min + kw \qquad (2)$$

Each feature that fall within the counterpart represented by each interval or bin is transformed into the same representation value.

The first representation value is calculated as below:

$$l = min \qquad u = min + w$$
$$r = (l + u)/2 \qquad (3)$$

This value, r represents the universal value or the common figure that is used to transform any features data value within the range of the first interval or bin, from l that acts as its first lower boundary value until u that presents the first upper boundary value. This process of calculating representation value is done iteratively based on the consequence lower and upper boundary values for each bin or interval until it reaches the number of features data, k that is also the same as the number of total bins or intervals. Any features data that falls within the consequence lower and upper boundary value is transformed using its own representation value respectively.

This study reconstructs the formulation of discretization method and implements them towards all features to enhance the performance of classifier models for author's handwriting image recognition.

III. INVARIANT DISCRETIZATION

A. Representation Values based Discretization

This study has implemented the Higher Order United Moment Invariant (HUMI) [16]–[17] as feature extraction method to produce features that represent the Global Features while the Edge based Directional (ED) [9] method is executed to extract the Local Features from the handwriting image for author's recognition. The result of discretization procedure is shown in Table 1 to produce the global discretized feature values for Writer 1 (W1) of HUMI. Besides, the same procedure of discretization is done for ED features to generate the Discretized Local Features shown by Table 2 of W1.

The discretization procedure is started by finding the minimum and maximum value throughout all feature values for W1. Both values are used to calculate the interval or bin value for the W1. Table 1 shows the minimum value of global features for W1 is 0.000613994 while the maximum value is 6.95071. The interval or bin value is calculated by finding the difference between the maximum and minimum value of W1. Next the value is divided by the number of features for HUMI that is eight (8). Based on the proposed formulation of (1), (2) and (3) in section Methodology, the first interval or bin value is calculated as 0.869376 for global features of W1. The minimum value is then being set as the first lower value and the first interval or bin value is set as the first upper value.

This is constructed as the range of the first bin named as Bin 0.

Table 1: Discretization Process to produce Representation Values of Global Features for Writer 1

Bin	Lower	Upper	Rep Value
0	0.000613994	0.869376	0.434995
1	0.869376	1.73814	1.30376
2	1.73814	2.6069	2.17252
3	2.6069	3.47566	3.04128
4	3.47566	4.34442	3.91004
5	4.34442	5.21319	4.7788
6	5.21319	6.08195	5.64757
7	6.08195	6.95071	6.51633

Table 2: Discretization Process to produce Representation Values of Local Features for Writer 1

Bin	Lower	Upper	Rep Value
0	0.7716	2.32636	1.54898
1	2.32636	3.88111	3.10373
2	3.88111	5.43587	4.65849
3	5.43587	6.99062	6.21324
4	6.99062	8.54538	7.768
5	8.54538	10.1001	9.32276
6	10.1001	11.6549	10.8775
7	11.6549	13.2096	12.4323
8	13.2096	14.7644	13.987

All the eight bins are constructed to determine the range for each bin. The second bin is determined by using the first upper value for Bin 0 as the lower value of Bin 1. Each bin value is constructed by adding up the interval or bin value to determine the lower and upper value. Both values are used to calculate the interval or bin value for the particular writer.

While Table 2 shows the minimum value of local features for W1 is 0.7716 and the maximum value is 14.7644. The first interval or bin value is calculated as 0.7716 for local features of W1. The minimum value is then set as the first lower value and the first interval or bin value is set as the first upper value. Next, a representation value that is labeled as *Rep Value* in Table 1 and Table 2 is calculated based on the lower and upper value to find the middle common value that is used to represent the features.

It is calculated by adding up the lower and upper value and divides them with two. Table 1 shows this *Rep Value*, 0.434995 for global features and 1.54898 for local features shown by Table 2 for Bin 0 respectively. The *Rep Value* is considered as the general or universal features that represents any value that fall between the range of the lower and upper value for the particular Bin. This procedure is repeated for the global and local features of all 30 writers to transform them into discretized feature vectors.

B. Discretized and Non-discretized Features

The Global Features in this study is produced by using the Higher Order United Moment Invariant (HUMI) [16]–[17] feature extraction method. Table 3 and Table 4 show the production of eight (8) non-discretized HUMI features for the Writer Class 1 and Writer Class 2. Besides, the Local Features are presented in Table 5 and Table 6 by using the Edge based Directional (ED) [9] feature extraction method that constructed nine (9) features for Writer Class 1 and Writer Class 2. Table 3 until 6 show only partial data of the whole word images per writer. There are in total thirty (30) authors with twenty (20) images per authors that given the total handwriting images used are 600 word images for the purpose of experimentation. Handwriting images are retrieved from the IAM Online Handwriting Database (IAM-OnDB) [21].

As shown by Table 3, HUMI has produced feature values that are rather high in similarity between writers and between features. For an example, the value of the first feature, F1 for Writer Class 1 is 4.21241 for the word image "By" while for Writer Class 2 is 4.25156 for the word image "General". Another example of the high similarity values between both writers has been given by the eighth feature, F8 that present the value of 0.506361 for Writer Class 1 and 0.532863 for Writer Class 2. The distributions among features for Writer Class 1 ranging from the minimum value of 0.000613994 until the maximum value of 6.95071 have shown that there is also high similarity among all eight features by considering the difference between each feature values. The high similarity between features is shown by F1 that has given the value of 4.17679 while the value of 4.178 for F2 extracted from the same word image "Trevor" for Writer Class 1.

Besides, the ED has constructed feature values that are also high in similarity between writers as shown by Table 5. For example, the value of the first feature, F1 for Writer Class 1 and Writer Class 2 is the same that is 0.97 for the word image "By" and "General". However, the local features (ED) have a big difference among the distribution of features starting with minimum values of 0.7716 and ending with the maximum values of 14.7644. This has given the ability to diverge among all the nine (9) Local Features (ED) that have a slightly lower similarity between features. This is shown by the difference between F1 value, 14.7 and F2, 1.0 that has given lower similarity between features.

For both feature extraction methods HUMI and ED, the cases of high similarity factor between writers and between features, the individuality for each writer is difficult to be determined by the classifier model. Thus, it will lead to the poor and lower identification performance.

Table 7 and 8 show the discretization feature vectors for Global Features (HUMI). In the table, feature F1 and F3 have shown through the dotted circle that there are two (2) most frequent discretized feature vectors; 4.7788 and 3.91004 for F1 while 0.434995 and 1.30376 for F3 that are used to represent and transform the features for Writer Class 1. Feature F2, F4 and F8 is represented by three (3) most frequent discretized feature vectors that are shown through the dotted circle while F5 is represented by four (4) most frequent. The rest of the Global Features (HUMI) is represented by at least five (5) most frequent discretized feature vectors. The discretized feature vectors constructed by Writer Class 1 show dissimilarity pattern to be compared with Writer Class 2 thus given the ability to distinguish the writers.

Besides, Table 9 and 10 present the discretization feature vectors for Local Features (ED). Feature F2, F3, F7, F8 and F9 have been discretized into only one most frequent discretized feature vector that is shown by the dotted circle. This feature vector is used in transforming the Local Features for Writer Class 1. The rest of the features include F1 that is represented by three (3), F4 and F6 are represented by four (4) and F5 is represented by five (5) most frequent discretized feature vectors. Both Global and Local Features have been infused with generalization factor that able to present the individuality and uniqueness of each writer. This has also shown that the local features have been able to be generalized by discretization that can be represented by only one feature vector for certain writer. Thus, it can then lead to better and high performance for classifier model.

Table 3: Partial Data of Non-Discretized Global Features for Writer 1

WORD	F1	F2	F3	F4	F5	F6	F7	F8
By	4.21241	4.38808	0.1132	0.3945	3.86873	4.55612	3.99357	0.506361
Trevor	4.17679	4.178	0.6220	1.2488	1.68269	6.67092	2.92916	1.87226
Williams	4.7278	5.84875	0.0914	0.1830	3.24089	6.21473	5.66567	0.430031
Move	4.61117	4.85619	0.2565	0.5333	3.3198	5.90255	4.32286	0.827221
To	4.56784	4.95431	0.1895	0.0120	4.54839	4.58731	4.94223	0.344743
Stop	4.21165	4.90193	0.0132	0.8372	2.71065	5.71269	4.06469	0.886424
From	4.08356	4.19513	0.6062	1.5431	1.21644	6.95071	2.65199	2.15011
nominating	4.61196	4.8902	0.2132	0.1884	4.57184	4.65208	4.70177	0.411361
Any	4.10236	4.12798	0.0452	0.1517	3.83452	4.37024	3.97622	0.393582
More	4.72282	4.72311	0.0402	0.6388	4.00319	5.44247	4.08428	0.754941

Table 4: Partial Data of Non-Discretized Global Features for Writer 2

WORD	F1	F2	F3	F4	F5	F6	F7	F8
General	4.25156	4.73691	0.3610	0.4081	4.08012	4.42303	4.32878	0.532863
official	4.46295	4.51434	0.2093	1.2091	3.62102	5.30489	3.30518	1.06505
Welcome	4.5699	4.76373	0.5481	0.0005	5.47295	3.66687	4.76427	0.581597
Last	3.73379	3.78177	0.5827	1.3225	1.19767	6.27	2.45918	1.90678
Week	4.36434	4.73998	0.2155	0.9779	3.44188	5.28682	3.76201	0.870928
To	4.13248	6.00411	0.6999	1.5825	0.721536	8.98653	4.4216	2.2829
Britain	4.31391	4.38593	0.1216	0.3021	3.69632	4.93152	4.08374	0.53269
Moves	4.18494	4.35809	0.0168	1.1517	2.89369	5.47623	3.2063	1.16684
Towards	4.04054	4.05429	0.2746	1.2256	2.25195	5.82917	2.82864	1.50748
The	4.51458	4.52102	0.4198	0.8397	2.82874	6.20043	3.68132	1.26854

Table 7: Partial Data of Discretized Global Features for Writer 1

WORD	F1	F2	F3	F4	F5	F6	F7	F8
By	3.91004	4.7788	0.434995	0.434995	3.91004	4.7788	3.91004	0.434
Trevor	3.91004	3.91004	0.434995	1.30376	1.30376	6.51633	3.04128	2.172
Williams	4.7788	5.64757	0.434995	0.434995	3.04128	6.51633	5.64757	0.434
Move	4.7788	4.7788	0.434995	0.434995	3.04128	5.64757	3.91004	0.434
To	4.7788	4.7788	0.434995	0.434995	4.7788	4.7788	4.7788	0.434
Stop	3.91004	4.7788	0.434995	0.434995	3.04128	5.64757	3.91004	1.303
From	3.91004	3.91004	0.434995	1.30376	1.30376	6.51633	3.04128	2.172
nominating	4.7788	4.7788	0.434995	0.434995	4.7788	4.7788	4.7788	0.434
Any	3.91004	3.91004	0.434995	0.434995	3.91004	4.7788	3.91004	0.434
More	4.7788	4.7788	0.434995	0.434995	3.91004	5.64757	3.91004	0.434

Table 8: Partial Data of Discretized Global Features for Writer 2

WORD	F1	F2	F3	F4	F5	F6	F7	F8
General	3.93191	5.05516	0.56216	0.56216	3.93191	3.93191	3.93191	0.562
official	3.93191	5.05516	0.56216	1.68541	3.93191	5.05516	2.80866	0.562
Welcome	5.05516	5.05516	0.56216	0.56216	5.05516	3.93191	5.05516	0.562
Last	3.93191	3.93191	0.56216	1.68541	1.68541	6.17841	2.80866	1.685
Week	3.93191	5.05516	0.56216	0.56216	3.93191	5.05516	3.93191	0.562
To	3.93191	6.17841	0.56216	1.68541	0.56216	8.42491	3.93191	2.808
Britain	3.93191	3.93191	0.56216	0.56216	3.93191	5.05516	3.93191	0.562
Moves	3.93191	3.93191	0.56216	1.68541	2.80866	5.05516	2.80866	1.685
Towards	3.93191	3.93191	0.56216	1.68541	2.80866	6.17841	2.80866	1.685
The	5.05516	5.05516	0.56216	0.56216	2.80866	6.17841	3.93191	1.685

Table 9: Partial Data of Discretized Local Features for Writer 1

WORD	F1	F2	F3	F4	F5	F6	F7	F8	F9
By	1.54898	1.5489	1.548	1.548	4.658	1.548	1.54898	1.54898	1.54898
Trevor	13.987	1.5489	1.548	1.548	1.548	1.548	1.54898	1.54898	1.54898
Williams	1.54898	1.5489	1.548	3.103	1.548	4.658	1.54898	1.54898	1.54898
Move	1.54898	1.5489	1.548	9.322	1.548	1.548	1.54898	1.54898	1.54898
To	1.54898	1.5489	1.548	1.548	1.548	1.548	1.54898	1.54898	1.54898
Stop	1.54898	1.5489	1.548	1.548	1.548	1.548	1.54898	1.54898	1.54898
From	1.54898	1.5489	1.548	1.548	1.548	1.548	1.54898	1.54898	1.54898
Nominating	1.54898	1.5489	1.548	1.548	6.213	1.548	1.54898	1.54898	1.54898
Any	1.54898	1.5489	1.548	1.548	1.548	1.548	1.54898	1.54898	1.54898
More	1.54898	1.5489	1.548	9.322	1.548	6.213	1.54898	1.54898	1.54898

Table 5: Partial Data of Non-Discretized Local Features for Writer 1

WORD	F1	F2	F3	F4	F5	F6	F7	F8	F9
By	0.97	0.98	0.9412	0.9135	4.5693	0.8685	0.9412	0.91	1.2749
Trevor	14.7	1	0.9412	0.8754	0.8902	0.8543	0.9412	0.9412	0.8858
Williams	0.99	0.98	0.9412	3.6073	1.0388	4.9459	0.9412	0.9412	0.8858
Move	1	1	0.9412	10.076	1.0285	0.7958	0.9412	0.9412	0.8858
To	0.97	0.99	0.9412	1.1371	1.0623	0.8685	0.9412	0.9412	0.8858
Stop	1	1	0.9412	0.8651	1.0123	0.8028	0.9412	0.9412	0.8824
From	1.17	1	0.9412	0.8235	7.1429	1.0559	0.9412	0.9412	0.8858
nominating	1	0.99	0.9412	1.4464	5.6314	1.1693	0.9412	0.9412	0.8789
Any	1	1	0.9412	0.8824	0.7843	1.0182	0.9412	0.9412	0.8512
More	1	1	0.9412	0.3051	0.8408	6.6914	0.9412	0.9412	0.8858

Table 6: Partial Data of Non-Discretized Local Features for Writer 2

WORD	F1	F2	F3	F4	F5	F6	F7	F8	F9
General	0.97	1	0.9412	1.0002	3.8089	0.7578	0.9412	0.9412	0.8858
official	15.6	1	0.9412	0.9692	7.2629	7.9276	0.9412	0.9377	0.8858
welcome	1	1	0.9412	7.6894	0.9624	6.1952	0.9412	0.9412	0.8858
Last	0.99	1	0.9343	0.8962	0.7889	0.9919	0.9412	0.9412	0.8858
week	1	1	1.1348	9.5191	0.9943	1.0123	0.9412	0.9412	0.8858
To	0.99	1	0.9412	1.0897	0.8235	0.8651	0.9412	0.9412	0.8858
Britain	1	1	0.9412	0.8304	4.5367	0.8201	0.9412	0.9412	0.8858
Moves	1	1	0.9412	0.8547	5.6472	1.0071	0.9412	0.9412	0.8858
Towards	1	1	0.9412	0.8235	1.0422	0.782	0.9412	0.9412	0.8651
The	0.98	1	0.9412	0.8512	0.8443	0.8547	0.9412	0.9412	0.8858

Table 10: Partial Data of Discretized Local Features for Writer 2

WORD	F1	F2	F3	F4	F5	F6	F7	F8	F9
General	1.58495	1.5849	1.584	1.584	3.229	1.584	1.58495	1.58495	1.58495
official	14.8193	1.5849	1.584	1.584	6.547	8.202	1.58495	1.58495	1.58495
welcome	1.58495	1.5849	1.584	8.202	1.584	6.547	1.58495	1.58495	1.58495
Last	1.58495	1.5849	1.584	1.584	1.584	1.584	1.58495	1.58495	1.58495
Week	1.58495	1.5849	1.584	9.856	1.584	1.584	1.58495	1.58495	1.58495
To	1.58495	1.5849	1.584	1.584	1.584	1.584	1.58495	1.58495	1.58495
Britain	1.58495	1.5849	1.584	1.584	4.893	1.584	1.58495	1.58495	1.58495
Moves	1.58495	1.5849	1.584	1.584	4.893	1.584	1.58495	1.58495	1.58495
Towards	1.58495	1.5849	1.584	1.584	1.584	1.584	1.58495	1.58495	1.58495
The	1.58495	1.5849	1.584	1.584	1.584	1.584	1.58495	1.58495	1.58495

IV. RESULT AND DISCUSSION

The comparison of identification performance is done between the Non-Discretized and Discretized features for Global and Local Features to show the importance of discretization method in deploying the generalization factor towards feature vectors.

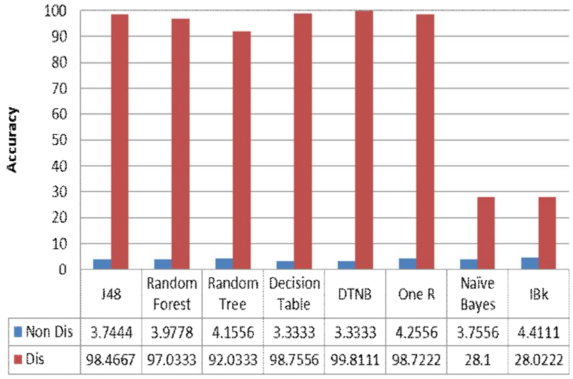
Global Features and Local Features for images of thirty (30) authors’ handwriting datasets are used to input eight (8) classifier schemes. The selected classifier schemes are the J48 Decision Tree, Random Forest, Random Tree, Decision Table, Decision Table Naïve Bayes (DTNB), One Attribute Rule (OneR), Naïve Bayes (NB) and the Instance based classifiers with k parameter (IBk) that is using the k nearest neighbor algorithm (k-NN). The experiments for these classifiers are set to perform five (5) types of training and testing environment settings for fifty (50) runs. The train and test runs on the authors’ datasets are implemented by using the five (5) and ten (10) fold cross validation, a setup of 60% of the patterns for training and 40% for testing, also the 70% of training and 30% of testing and lastly 80% for training and 20% for testing the datasets for all eight classifier schemes.

One of the performance results is shown by Fig. 1 with a setup of 70% for training and 30% for testing that shows the classifier DTNB has given the highest accuracy of 99.81% for all discretized features while non-discretized features only manage to get 3.33% accuracy rate from the same scheme. The highest accuracy rate for the non-discretized features is 4.41% given by IBk scheme which far worst to be compared with features that go through discretization procedure. As a result, discretized features for almost all classifiers are really at high performance which given the result above 90% presenting the second highest rate is Decision Table with 98.76%, OneR gets 98.72%, J48 produces 98.47%, Random Forest gives 97.03% and Random Tree with 92.03% excluding the classifiers of Naïve Bayes and IBk that are showing a rather lower performance at 28.1% and 28.02% respectively.

Fig. 2 shows the average comparison performance of every classifier schemes for all discretize and non-discretize features for five (5) experimental analysis setup. This is followed by the schemes of Decision Table, One R and J48 with the performance above 98%.

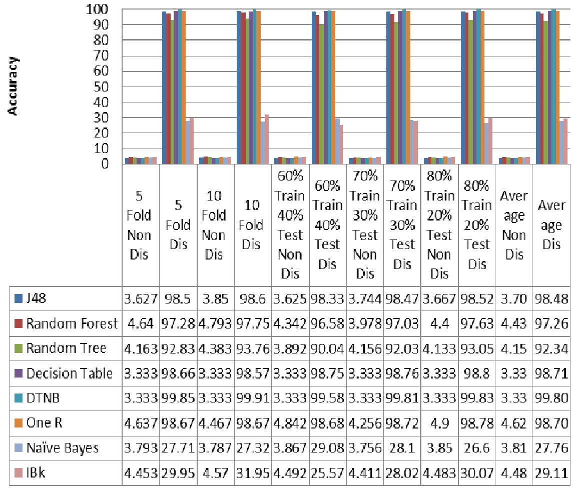
Besides, Random Forest scheme has achieved 97.26% in average while Random Tree gets 92.34% accuracy rate. Though, two schemes have achieved quite a lower average performance of 29.11% for IBk and 27.76% for Naïve Bayes but both have presenting higher and better performance values than all the classifier schemes for all non-discretized features that only able to perform in average up until 4.62%.

This achievement has proposed that discretization procedure for HUMI features has contributed to the higher performance rate by representing the features with generalization factor. Consecutively, Fig. 3 shows the performance of discrete features and non-discretize features for Local Features. The Classifier DTNB has presented the best performance accuracy of 99.89% for all discretized features while non-discretized features only manage to get 3.33% accuracy rate for the 70% training and 30% testing environment setup.



	J48	Random Forest	Random Tree	Decision Table	DTNB	One R	Naïve Bayes	IBk
🟦 Non Dis	3.7444	3.9778	4.1556	3.3333	3.3333	4.2556	3.7556	4.4111
🟥 Dis	98.4667	97.0333	92.0333	98.7556	99.8111	98.7222	28.1	28.0222

Fig. 1: Performance of All Discretize and Non-Discretized Global Features for Split Percentage of 70% Training and 30% of Testing with fifty (50) runs

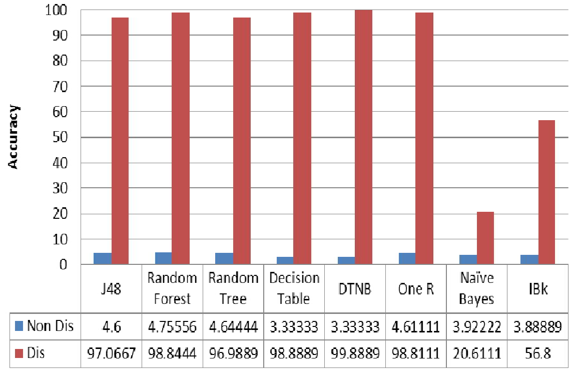


	5 Fold Non Dis	5 Fold Dis	10 Fold Non Dis	10 Fold Dis	60% Train 40% Test Non Dis	60% Train 40% Test Dis	70% Train 30% Test Non Dis	70% Train 30% Test Dis	80% Train 20% Test Non Dis	80% Train 20% Test Dis	Average Non Dis	Average Dis
J48	3.627	98.5	3.85	98.6	3.625	98.33	3.744	98.47	3.667	98.52	3.70	98.48
Random Forest	4.64	97.28	4.793	97.75	4.342	96.58	3.978	97.03	4.4	97.63	4.43	97.26
Random Tree	4.163	92.83	4.383	93.76	3.892	90.04	4.156	92.03	4.133	93.05	4.15	92.34
Decision Table	3.333	98.66	3.333	98.57	3.333	98.75	3.333	98.76	3.333	98.8	3.33	98.71
DTNB	3.333	99.85	3.333	99.91	3.333	99.58	3.333	99.81	3.333	99.83	3.33	99.80
One R	4.637	98.67	4.467	98.67	4.842	98.68	4.256	98.72	4.9	98.78	4.62	98.70
Naïve Bayes	3.793	27.71	3.787	27.32	3.867	29.08	3.756	28.1	3.85	26.6	3.81	27.76
IBk	4.453	29.95	4.57	31.95	4.492	25.57	4.411	28.02	4.483	30.07	4.48	29.11

Fig. 2: Average Comparison Performance of All Discretize and Non-discretized Global Features for Different Experimental Analysis Setup

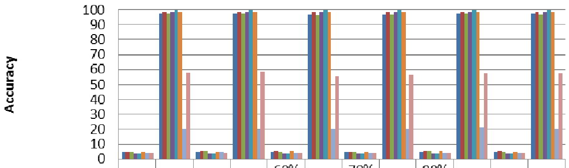
The performance for all non-discretized features for all schemes and environment setup can only reach as high as 4.76% for the classifier of Random Forest. In comparison with all discretized features for another three (3) schemes that include the Random Forest, Decision Table and OneR have performed higher than 98% of classification accuracy rate. J48 performs at 97.07% while 96.99% is achieved by Random Tree. A slightly lower performance is presented by IBk at 56.8% and the lowest is performed by Naiive Bayes at 20.61%.

The performance of DTNB in average for discretize features has succeeded as the best performance for all classifier schemes that reached the highest of 99.89% accuracy rate. This is shown by Fig. 4. The second highest includes three classifiers that are the Random Forest, Decision Table and One R have presented the performance of more than 98% in average. Besides, J48 has given 97.18% while Random Tree has performed 97% for all discretized features. However, there are two classifier scheme that performed quite lower as compared to the others with discretize features.



	J48	Random Forest	Random Tree	Decision Table	DTNB	One R	Naïve Bayes	IBk
🟦 Non Dis	4.6	4.75556	4.64444	3.33333	3.33333	4.61111	3.92222	3.88889
🟥 Dis	97.0667	98.8444	96.9889	98.8889	99.8889	98.8111	20.6111	56.8

Fig. 3: Performance of All Discretize and Non-Discretized Local Features for Split Percentage of 70% Training and 30% of Testing with fifty (50) runs



	5 Fold Non Dis	5 Fold Dis	10 Fold Non Dis	10 Fold Dis	60% Train 40% Test Non Dis	60% Train 40% Test Dis	70% Train 30% Test Non Dis	70% Train 30% Test Dis	80% Train 20% Test Non Dis	80% Train 20% Test Dis	Average Non Dis	Average Dis
■ J48	4.65	97.26	4.797	97.43	4.633	97.04	4.6	97.07	4.717	97.13	4.679	97.18
■ Random Forest	4.69	98.92	5.05	98.9	4.9	98.86	4.756	98.84	4.983	98.87	4.875	98.87
■ Random Tree	4.767	97.11	4.84	97.27	4.658	96.53	4.644	96.99	4.833	97.15	4.748	97.00
■ Decision Table	3.333	98.93	3.333	99.06	3.333	98.85	3.333	98.89	3.333	98.7	3.333	98.88
■ DTNB	3.333	99.93	3.333	99.95	3.333	99.81	3.333	99.89	3.333	99.92	3.333	99.89
■ One R	4.797	98.86	4.63	99.06	4.883	98.91	4.611	98.81	4.933	98.62	4.770	98.85
■ Naïve Bayes	4.037	20.57	4.363	20.11	4.15	20.23	3.922	20.61	4.217	20.8	4.137	20.46
■ IBk	3.867	57.46	3.907	58.42	4	55.48	3.889	56.8	3.983	56.97	3.929	57.02

Fig. 4: Average Comparison Performance of All Discretize and Non-discretized Local Features for Different Experimental Analysis Setup

The classifier IBk has presented the performance accuracy of 57.02% while Naïve Bayes can only reach as high as 20.46% in average. Although, both performances are still substantially higher to be compared with the highest performance for all non-discretized features that can only performed in average up until 4.88% for classifier Random Forest. This has shown that discretization method has been able to implement the generalization factor into Local Features and significantly increased their classification performance.

V. CONCLUSION

As a conclusion, the comparison performance for all discretized and non-discretize features have shown the importance of discretization method in this study. The Local Features has performed the best in average at 99.89% from the classifier DTNB while Global Features performed at 99.81%. Without discretization procedures, Global Features can only reach the performance of 4.62% in average while Local Features has managed to go a little bit higher for 4.88% in average performance. This is a very poor performance in comparison towards the performance of discretized features. This is due to the high similarity for feature values between writers and between features for both feature extraction methods as discussed earlier. The high similarity between features has caused a problem for the classifiers to distinguish between features and differentiate the writers' characteristics for the handwriting image. This has led to the poor identification performance. Thus, the proposed discretization method has shown such promising results for the handwriting image analysis for authors' recognition. The generalization factor infused by discretization method has been able to boost the performance for both Global and Local Features.

ACKNOWLEDGEMENT

The research is supported by Universiti Teknikal Malaysia Melaka (UTeM). The research has been carried out under the Fundamental Research Grant Scheme (FRGS) project FRGS/2018/FTMK-CACT/F00390 provided by Ministry of Education of Malaysia. This research is also supported by Computational Intelligence and Technologies Lab (CIT Lab) research group under the Centre for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat Dan Komunikasi (FTMK), UTeM.

REFERENCES

[1] V. N. Temlyakov, "*Universal discretization*," Journal of Complexity, 47, 2018, 97-109.

[2] E. Hirvijoki, E & M. F. Adams, "*Conservative discretization of the Landau collision integral*," Physics of Plasmas, 24(3), 2017, 032121.

[3] X. Zhao, P. Shi, & X. Zheng, "*Fuzzy adaptive control design and discretization for a class of nonlinear uncertain systems*," IEEE transactions on cybernetics, 46(6), 2015, 1476-1483.

[4] A. C. Luo, "*Discretization and implicit mapping dynamics*," Springer Berlin Heidelberg, 2015

[5] I. E. A. Jalil, S. M. Shamsuddin, A. K. Muda and A. Ralescu, "*Geometrical feature based ranking using grey relational analysis (GRA) for writer identification*," 2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR), 2013, pp 152-157 IEEE

[6] I. E. A. Jalil, S. M. Shamsuddin, A. K. Muda, M. S. Azmi and U. R. Hashim, "*Predictive based hybrid ranker to yield significant features in writer identification*," International Journal of Advances in Soft Computing & Its Applications 10(1), 2018

[7] A. Ali & B. Omer, "*Invarianceness for Character Recognition Using Geo-Discretization Features*," Computer and Information Science, 9(2), 2016, 1.

[8] J. Biesiada, W. Duch & A. Kachel, "*Feature ranking methods based on information entropy with Parzen windows*," 2005

[9] W. Y. Leng & S. M. Shamsuddin, "*Writer Identification for Chinese Handwriting*," 2(2), 2016.

[10] A. K. Muda, S. M. Shamsuddin & M. Darus, "*Discretization of integrated moment invariants for writer identification*," In Proceeding of The 4th IASTED International Conference on Advances in Computer Science and Technology, 2008, (pp. 372-377).

[11] A. K. Muda, S. M. Shamsuddin, & M. Darus, " *Invariants Discretization for Individuality*", 2008, 218–228.

[12] T. Winiarski, J. Biesiada, A. Kachel, W. Duch, T. Winiarski, J. Biesiada, J & A. Kachel, "*Feature ranking, selection and discretization*," In Proceedings of Int. Conf. on Artificial Neural Networks (ICANN), 2003, 251–254.

[13] B. O. Mohammed & S. M. Shamsuddin, "*Improvement in twins handwriting identification with invariants discretization*," EURASIP Journal on Advances in Signal Processing, 2012(1), 48.

[14] M. Dash & H. Liu, "*Feature Selection for Clustering*," Knowledge Discovery and Data Mining. Current Issues and New Applications, 1805(8), 2000, 110–121.

[15] S. Kotsiantis & D. Kanellopoulos, "*Discretization techniques: A recent survey*," GESTS International Transactions on Computer Science and Engineering, 32(1), 2006, 47-58.

[16] A. K. Muda, S. M. Shamsuddin & A. Abraham, "*Authorship Invarianceness for Writer Identification*," 2009, 34–39.

[17] S. M. Shamsuddin, M. N. Sulaiman, & M. Darus, "*Invarianceness of Higher Order Centralised Scaled-invariants Undergo Basic Transformations*," International Journal of Computer Mathematics, 79(1), 2002, 39–48.

[18] T. C. Eng, S. Hasan, S. M. Shamsuddin, N. E. Wong and I. E. A. Jalil, "*Big data processing model for authorship identification*," International Journal of Advances in Soft Computing & Its Applications 9(3), 2017

[19] A. R. Radzid, M. S. Azmi, I. E. A. Jalil, N. A. Arbain, A. K. Draman, & A. Tahir (2018). "*Text line segmentation for mushaf Al-Quran using hybrid projection based neighbouring properties*." Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(2-7), 53-57.

[20] A. R. Radzid, M. S. Azmi, I. E. A. Jalil, A. K. Muda, L. B. Melhem, & N. A. Arbain, (2018). "*Framework of page segmentation for mushaf Al-Quran based on multiphase level segmentation*." Int. J. Comput. Inf. Syst. Ind. Manag. Appl., 10, 028-037.

[21] M. Liwicki & H. Bunke, (2007). "*Handwriting recognition of whiteboard notes—studying the influence of training set size and type*". International Journal of Pattern Recognition and Artificial Intelligence, 21(01), 83-98.