IMPROVEMENT OF AUTHORSHIP INVARIANCENESS FOR INDIVIDUALITY REPRESENTATION IN WRITER IDENTIFICATION

Azah Kamilah Muda[*], *Siti Mariyam Shamsuddin*[†], *Ajith Abraham*[‡]

Abstract: Writer Identification (WI) is one of the areas in pattern recognition that have created a center of attention for many researchers to work in. Recently, its main focus is in forensics and biometric application, e.g. writing style can be used as biometric features for authenticating individuality uniqueness. Existing works in WI concentrate on feature extraction and classification task in order to identify the handwritten authorship. However, additional steps need to be performed in order to have a better representation of input prior to the classification task. Features extracted from the feature extraction task for a writer are in various representations, which degrades the classification performance. This paper will discuss this additional process that can transform the various representations into a better representation of individual features for Individuality of Handwriting, in order to improve the performance of identification in WI.

Key words: *Writer identification, individuality representation, authorship invarianceness, discerization*

Received: March 20, 2009
Revised and accepted: February 10, 2010

1. Introduction

In the development of the digital age, paper documents are still exchanged. In some situations, Writer Identification (WI) is needed to identify the owner of handwriting. Handwriting identification can be included as a particular kind of dynamic biometric where the shapes and writing styles of writing can be used as biometric features for authenticating an identity [1–4]. Typically, WI is performed on

[*] Azah Kamilah Muda
Universiti Teknikal Malaysia Melaka, 75450 Melaka, Malaysia, E-mail: azah@utem.edu.my

[†] Siti Mariyam Shamsuddin
Universiti Teknologi Malaysia, 81310 Johor, Malaysia, E-mail: mariyam@utm.my

[‡] Norwegian University of Science and Technology, Trondheim, Norway, E-mail: ajith.abraham@ieee.org

legal papers by means of signature. However, it can also be necessary to recognize the handwritten authorship without signature, such as in case of threatening letter, authorship determination of an old or historical manuscript. The writer can be identified by using limited handwritten text from handwriting. It has a great importance in the criminal justice system and it is widely explored in forensic handwriting analysis [5–8]. There are many issues and scenarios in WI still posing challenges which require further investigations and explorations.

Existing works in WI concentrate on feature extraction and classification task in order to identify the handwritten authorship. However, additional steps need to be performed in order to have a better representation of input prior to the classification task. Features extracted from the feature extraction task for a writer are in various representations, which degrades the identification performance. This paper will discuss an additional process used for transforming the various representations resulting from the feature extraction task into a better representation of individual features. The remainder of the paper is structured as follows. In Section 2, an overview of handwriting individuality is presented. The needs of the proposed process for individuality representation and a brief review of discretization are provided in Section 3 and Section 4, respectively. In section 5, detailed explanations of the proposed solution in the WI domain follow. The experiments to prove the proposed process are discussed in Section 6. Finally, conclusion is drawn in Section 7.

2. Individuality of Handwriting

Handwriting has long been considered as individualistic. It rests on the hypothesis that each individual has consistent handwriting [1, 9–14]. Fig. 1 shows the handwriting of the same words and Fig. 2 of different words by three writers. Each person is seen as having a specific texture [9, 14] and can be observed in the both figures. The shape is slightly different for the same writer and quite different for different writers. This is known as Individuality of Handwriting. It can be measured by similarity measurement of variance between features of a writer (intra-class), which must be lower than different writers (inter-class) [19, 13, 15, 16]. Good individual features must obtain the lowest similarity error for intra-class and the highest similarity error for inter-class. Therefore, it is vital to acquire individual features from handwriting to satisfy this requirement for identifying the handwritten authorship. The individuality of handwriting concept is defined as authorship invarianceness, which has been discussed in [17].

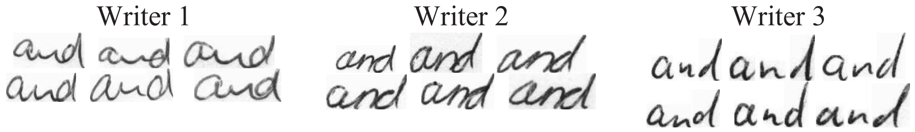


Fig. 1 *Same word by different writers.*

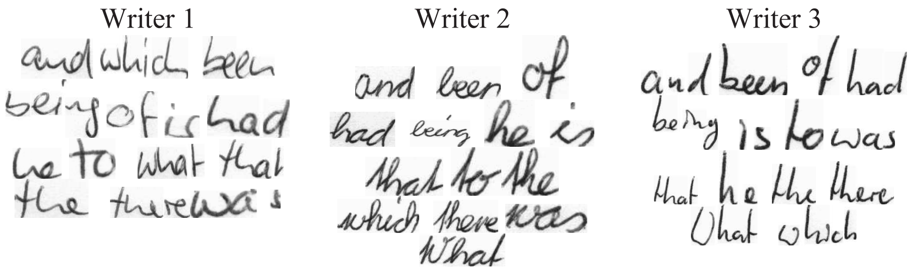


Fig. 2 *Different word by different writers.*

3. Individuality Representation

Good features as input to a classifier are important in order to obtain good performance in identification. Usually, extracted features directly perform the classification task in order to identify a writer. These features do not portray individual features of a writer, because the writer is represented by various features. Various representations lead to a large variance between features for intra-class (same writer) and small variance in inter-class (different writers). Apparently, another process is needed to improve the authorship invarianceness. This proposed process is known as invariant discretization and it is meant to reduce the variance between features for intra-class and increase the variance between features for inter-class. An overview of investigation that leads to the need of an additional procedure prior to the classification task in order to improve the identification performance of WI is shown in Fig. 3 below.

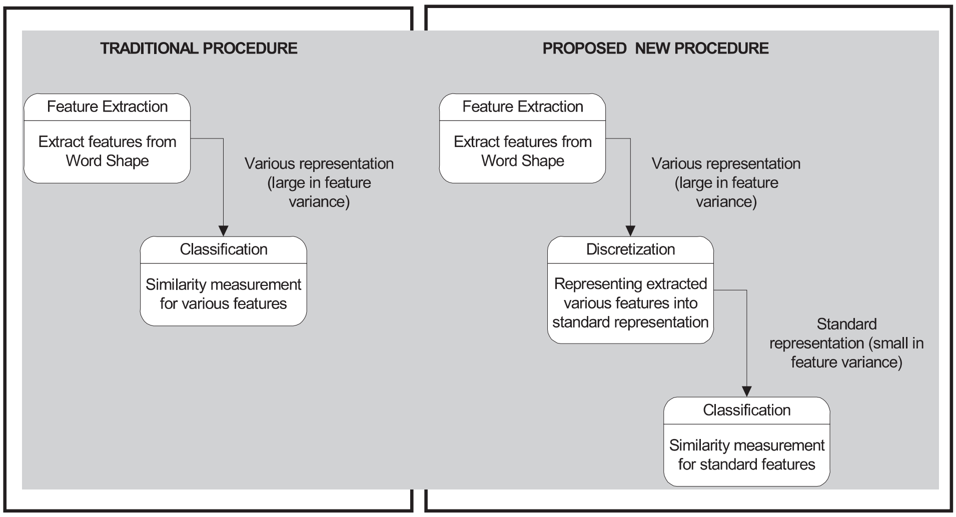

TRADITIONAL PROCEDURE
Feature Extraction
Extract features from Word Shape
Various representation (large in feature variance)
Classification
Similarity measurement for various features
PROPOSED NEW PROCEDURE
Feature Extraction
Extract features from Word Shape
Various representation (large in feature variance)
Discretization
Representing extracted various features into standard representation
Standard representation (small in feature variance)
Classification
Similarity measurement for standard features

Fig. 3 *An overview of the investigation.*

4. Discretization

Discretization is a process of dividing a range of continuous attributes into disjoint regions (interval), whose labels can then be used for replacing the actual data values [18]. It involves searching for "cuts" that determine intervals and unifying the values over each interval. All values that lie within an interval are mapped to the same value, in effect, converting numerical attributes that can be treated as symbolic [19]. Discretization is claimed as a solution to the problem of rough set theory, which cannot deal with the continuous attributes and a very large proportion of real data sets including continuous variables [20]. As reported in [21], through the discretization process, data sets are closer to a knowledge representation, they are easier to understand, use, explain, and the data can also be reduced and simplified. Moreover, the learning process becomes more accurate and faster. However, empirical results show the superiority of classification methods depends on the discretization algorithm used in the preprocessing process.

There are many discretization algorithms, which can be categorized in three basic perspectives. They are supervised versus unsupervised, global versus local, and dynamic versus static [22]. In the supervised method, class information is in hand, while no classification information is available for the unsupervised method. The perspective of global versus local describes the global method as discretizing all attributes in entire space while the local method discretizes a specific attribute of defined data [23]. The global discretization also performs preprocessing prior to the process of constructing a classifier, while the local method performs discretization during the process of classification [18]. Furthermore, the static versus dynamic perspective explains that the static method discretizes each attribute independently without considering interaction between attributes. On the other hand, the dynamic method consideres attributes interdependencies in the discretization process.

The supervised discretization is better than the unsupervised one due to its methodology of keeping the class information in the discretization process. The supervised discretization is introduced since the unsupervised method has a problem in dividing all instances into sub ranges by user specified width (range of values) or frequency (number of instances in each interval), which later leads to a poor result because the distribution of the continuous values is not uniform. In the supervised method, class information is used for finding the proper and meaningful intervals caused by cuts-points [21]. The work of Kerber in Dougherty [22] mentions that since the unsupervised method does not utilize instance labels in setting partition boundaries, classification information is lost in the binning process, which results from combining values that strongly associate with different classes into the same bin. The supervised method is also seen as coherent with the Individuality of Handwriting concept in the WI domain because individual features are preserved for the writer class.

5. Adaptation in Writer Identification

This section describes adaptations of the proposed solution in illustrating individual features and improving the variance between features for intra-class and inter-class in WI. It consists of three phases as depicted in Fig. 1.

5.1 Feature extraction

Moment Function (MF) has been used in diverse fields ranging from mechanics and statistics to pattern recognition and image understanding [24] for feature extraction. Extensive usage of moments in image analysis and pattern recognition was inspired by Hu [25] and Alt [26]. MF is used to extract the global shape of image. However, extracting features that represent and describe the shape precisely is a difficult task. A good shape descriptor should be able to find a perceptually similar shape that undergoes basic transformation, i.e., rotated, translated, scaled and affined transformed shapes. Due to the weaknesses of Hu's invariants, Yinan [27] proposed United Moment Invariant (UMI), where the rotation, translation and scaling can be discretely kept invariant to *region*, *closed* and *unclosed boundary.* It provides a good set of discriminated shape features and is valid in the discrete condition. It has also been derived to relate with a geometrical representation of Geometric Moment Invariant (GMI) [25] by considering the normalized central moments as:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\frac{p+q+2}{2}}}\mathrm{s}, \tag{1}$$

where $p + q = 2, 3, ...$; and in the discrete form, it is given as:

$$\begin{aligned} \eta'_{pq} &= \rho^{p+q}\eta_{pq} \\ &= \frac{\rho^{p+q}}{\mu_{00}^{\frac{p+q+2}{2}}}\mu_{pq}. \end{aligned} \tag{2}$$

and the Improved Moment Invariant (IMI) by Chen [28] with the given equation:

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

Equation (1), Equation (2) and Equation (3) have the factor μ_{pq}. By ignoring the influence of μ_{00} and ρ, the UMI [27] is given as:

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

where ϕ_i are Hu's moment invariants.

5.2 Proposed invariant discretization

The proposed invariant discretization resembles the simplest unsupervised methods of Equal Width Binning. However, the proposed method is categorized as the

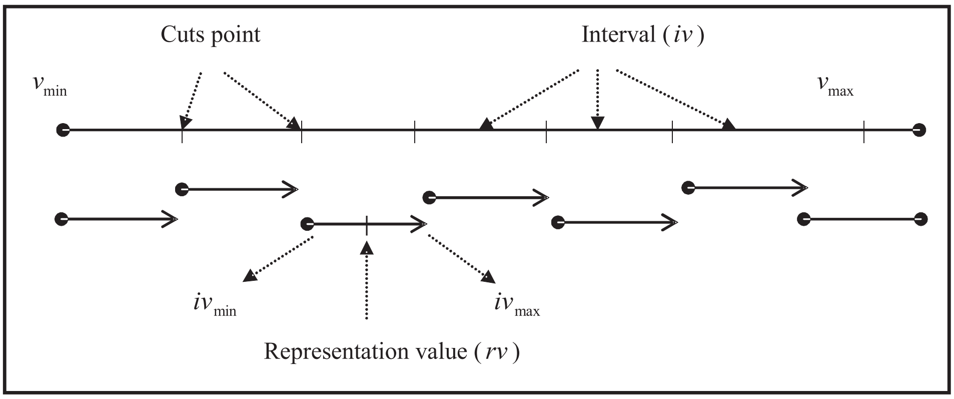


Fig. 4 *Invariant Discretization Line.*

supervised method because class information is used in the discretization process. It globally processes all invariant feature vectors with a dynamic characteristic. It is defined as global discretization with a dynamic characteristic because the discretized process is performed prior to the identification task and depends on all attributes in the data set when obtaining the representation value of an interval. These three factors directly contributed to the invariant discretization process in order to improve the authorship invarianceness in the WI domain.

5.2.1 Invariant discretization process

A suitable set of interval to represent the extracted features with a representation value is calculated in the discretization process. This representation value is called a discretized feature vector, where the "generalized unique feature" of individual features is obtained from the median of an interval. This generalized feature is used to illustrate the individual feature that is hidden in the individual writing of a writer. To acquire an interval, the range of minimum and maximum data of each writer is divided into a number of interval (cuts) with an equal size. The number of interval is defined based on the number of feature vector columns in the extracted features. As in the example, eight feature vector columns are obtained from the UMI technique. Lower and upper approximation is given to the each interval, and each of the intervals is represented by one representation value. The invariant feature vector that falls within the same interval will have the same representation. The representation value for an interval is calculated based on the writer class (supervised discretization). If two writers have a close or similar invariant feature vector, they will have the same or quite similar interval (cuts) for these two classes. The proposed algorithm does not change the information or characteristic of writer. It just represents the original extracted invariant feature vector in a standard representation with generalized features. Fig. 4 below provides an illustration of the invariant discretization line for the proposed discretization process.

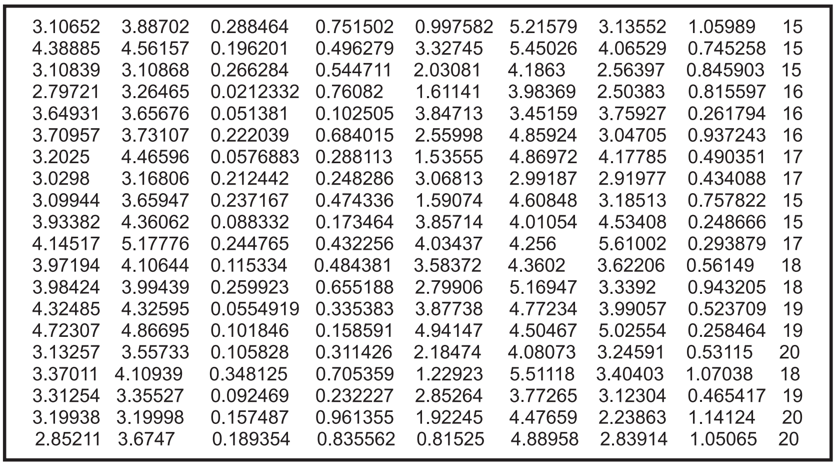
As mentioned above, the proposed invariant discretization needs the information of writer class in the discretization process. The minimum ($v_{\min}$) and the maximum ($v_{\max}$) invariant feature vectors (ifv) for a writer are used to calculate the range of intervals in the invariant discretization line. The line starts from the minimum ($v_{\min}$) invariant feature vector value and ends with the maximum ($v_{\max}$) invariant feature vector value for a particular writer. An interval is an average of invariant discretization line divided by the number of invariant feature vector column. The width (wd) of an interval can be calculated with:

$$wd = (v_{\max} - v_{\min})/f, \tag{4}$$

where :

$v_{\min}$: minimum value of invariant feature vector for a writer.
$v_{\max}$: maximum value of invariant feature vector for a writer.
f: the number of invariant feature vector column.

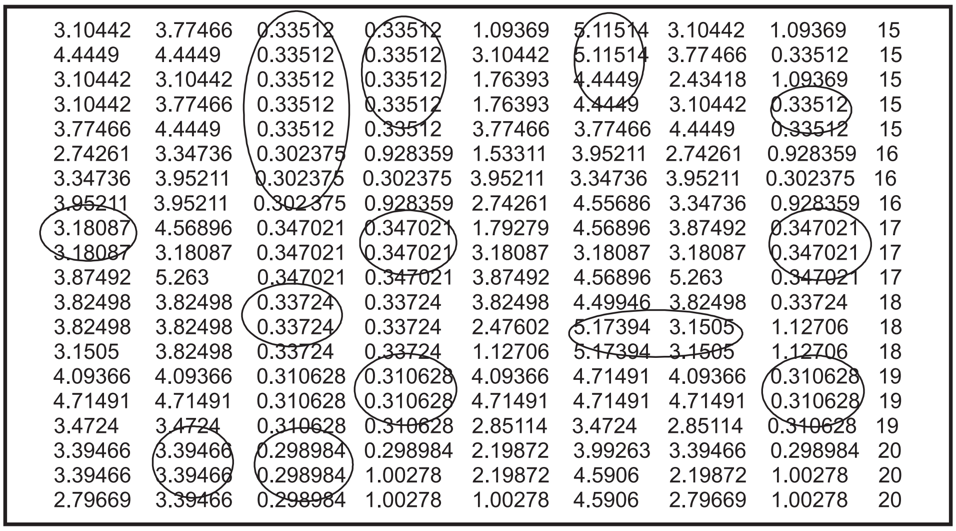
The width is used to define cut points of an interval in the invariant discretization line. Each invariant feature vector that falls within the same interval has the same representation value. The representation value (rv) for each interval is an average of interval that is calculated using$rv = (iv_{\max} - iv_{\min})/2$. The representation value for interval one to seven represents the invariant feature vector that falls within $ifv \geq iv_{\min}$ and $ifv < iv_{\max}$. Meanwhile, the invariant feature vector falls within $ifv \geq iv_{\min}$ and $ifv \leq iv_{\max}$ is categorized under the last interval. This representation value is known as discretized feature vector that represents the individual features of a writer. An example of transformation of invariant feature vector into a discretized feature vector is presented in Fig. 5 and Fig. 6, respectively.



3.10652	3.88702	0.288464	0.751502	0.997582	5.21579	3.13552	1.05989	15
4.38885	4.56157	0.196201	0.496279	3.32745	5.45026	4.06529	0.745258	15
3.10839	3.10868	0.266284	0.544711	2.03081	4.1863	2.56397	0.845903	15
2.79721	3.26465	0.0212332	0.76082	1.61141	3.98369	2.50383	0.815597	16
3.64931	3.65676	0.051381	0.102505	3.84713	3.45159	3.75927	0.261794	16
3.70957	3.73107	0.222039	0.684015	2.55998	4.85924	3.04705	0.937243	16
3.2025	4.46596	0.0576883	0.288113	1.53555	4.86972	4.17785	0.490351	17
3.0298	3.16806	0.212442	0.248286	3.06813	2.99187	2.91977	0.434088	17
3.09944	3.65947	0.237167	0.474336	1.59074	4.60848	3.18513	0.757822	15
3.93382	4.36062	0.088332	0.173464	3.85714	4.01054	4.53408	0.248666	15
4.14517	5.17776	0.244765	0.432256	4.03437	4.256	5.61002	0.293879	17
3.97194	4.10644	0.115334	0.484381	3.58372	4.3602	3.62206	0.56149	18
3.98424	3.99439	0.259923	0.655188	2.79906	5.16947	3.3392	0.943205	18
4.32485	4.32595	0.0554919	0.335383	3.87738	4.77234	3.99057	0.523709	19
4.72307	4.86695	0.101846	0.158591	4.94147	4.50467	5.02554	0.258464	19
3.13257	3.55733	0.105828	0.311426	2.18474	4.08073	3.24591	0.53115	20
3.37011	4.10939	0.348125	0.705359	1.22923	5.51118	3.40403	1.07038	18
3.31254	3.35527	0.092469	0.232227	2.85264	3.77265	3.12304	0.465417	19
3.19938	3.19998	0.157487	0.961355	1.92245	4.47659	2.23863	1.14124	20
2.85211	3.6747	0.189354	0.835562	0.81525	4.88958	2.83914	1.05065	20

Fig. 5 *Example of Invariant Feature Vector.*

The discretized feature vector obtained from the invariant discretization clearly illustrates the individual features of each writer.



3.10442	3.77466	0.33512	0.33512	1.09369	5.11514	3.10442	1.09369	15
4.4449	4.4449	0.33512	0.33512	3.10442	5.11514	3.77466	0.33512	15
3.10442	3.10442	0.33512	0.33512	1.76393	4.4449	2.43418	1.09369	15
3.10442	3.77466	0.33512	0.33512	1.76393	4.4449	3.10442	0.33512	15
3.77466	4.4449	0.33512	0.33512	3.77466	3.77466	4.4449	0.33512	15
2.74261	3.34736	0.302375	0.928359	1.53311	3.95211	2.74261	0.928359	16
3.34736	3.95211	0.302375	0.302375	3.95211	3.34736	3.95211	0.302375	16
3.95211	3.95211	0.302375	0.928359	2.74261	4.55686	3.34736	0.928359	16
3.18087	4.56896	0.347021	0.347021	1.79279	4.56896	3.87492	0.347021	17
3.18087	3.18087	0.347021	0.347021	3.18087	3.18087	3.18087	0.347021	17
3.87492	5.263	0.347021	0.347021	3.87492	4.56896	5.263	0.347021	17
3.82498	3.82498	0.33724	0.33724	3.82498	4.49946	3.82498	0.33724	18
3.82498	3.82498	0.33724	0.33724	2.47602	5.17394	3.1505	1.12706	18
3.1505	3.82498	0.33724	0.33724	1.12706	5.17394	3.1505	1.12706	18
4.09366	4.09366	0.310628	0.310628	4.09366	4.71491	4.09366	0.310628	19
4.71491	4.71491	0.310628	0.310628	4.71491	4.71491	4.71491	0.310628	19
3.4724	3.4724	0.310628	0.310628	2.85114	3.4724	2.85114	0.310628	19
3.39466	3.39466	0.298984	0.298984	2.19872	3.99263	3.39466	0.298984	20
3.39466	3.39466	0.298984	1.00278	2.19872	4.5906	2.19872	1.00278	20
2.79669	3.39466	0.298984	1.00278	1.00278	4.5906	2.79669	1.00278	20

Fig. 6 *Example of Descritized Feature Vector.*

5.3 Classification of identification task

Classification process assigns the digitized image to its symbolic class and classifies the image into categories based on the image data samples that classifier learned in its training stage [29]. The classification task is important in the WI domain because the extracted features must be classified accordingly in order to determine the owner of handwriting. The extracted feature must be a good set of features which represents the Individuality of Handwriting so that the handwritten authorship is reflected. Many approaches have been proposed for proving the Individuality of Handwriting concept in WI. It is to validate the capability of a technique in extracting individual features before it can be applied in the WI. From the literature, the approaches can be categorized as below:

a) Identification accuracy:

This approach is widely applied in the WI domain including [3, 10, 13, 30, 31, 32, 33, 34, 35, 36]. Many functions have been used as a classifier, such as Euclidean Distance, Hamming Distance, Nearest Neighbour classifier, Correlation measurer, Bayesian decision, Neural Network, SVM, Hidden Markov Models, etc. However, it is much better if the particular technique can be validated before the identification task because it can minimize the identification error of the identification task. Only the technique that satisfies the Individuality of Handwriting can be applied in the WI domain.

b) Clustering:

The clustering technique has been applied in [11, 12, 16, 37, 38]. It can be measured by features from the same writer grouped in the same cluster, and features from different writers in different clusters. There are many approaches that have been used to discriminate between features in the cluster, such as the similarity measurement of Euclidean distance, Hamming distance, Minkowski, Bhattacharya, the Kohonen self-organizing map (KSOM) and k-means clustering.

c) Similarity Measurement:

Cha and Srihari [39] proposed several similarity measurers in order to validate the distinctness of features between different writers. These are: convex hull of points, histogram distance, string distance of Levenshtein distance, and Euclidean distance. Manhattan distance has been applied in [40], together with genetic algorithm, in order to select a relevant feature that consists of individual features. Meanwhile, Brink *et al.*, [41] has created vantage profiles of a writer with Euclidean distance measurement in order to discriminate the writer with dissimilarity representations.

d) Statistical Functional:

Literature shows that statistical function also has been applied in many works in order to evaluate the individual features in the data set, such as standard deviation [42, 43] and mean square error [44].

Based on the literature review, it shows that the similarity measurement is mostly applied in classification, clustering or similarity measurement itself. This is due to the correspondence of similarity measurement to individuality of handwriting concept. Moreover, it is easy to implement. However, the identification task is only briefly reviewed in this paper because it focuses on the proposed invariant discretization process and not the classification task. The results of identification task are only shown in the next section as a proof that the proposed discretization can improve the identification performance in the WI domain.

6. Simulation Result

Two types of experiments have been conducted in this paper – authorship invarianceness and identification accuracy. First experiment is to prove the proposed invariant discretization can improve the variance between features for intra-class and inter-class. The latter experiment should evaluate the proposed discretization in improving the performance of identification using the Rosseta Toolkit [45]. In these experiments, IAM database [46] with 4 400 various images from 60 writers was used.

6.1 Authorship invarianceness

Authorship invarianceness is measured by using the Mean Absolute Error (MAE) function. Example of the MAE calculation is presented in Tab. I. The number of images is 20 for one author. Feature 1 to Feature 8 is an extracted feature that represents a word. The invarianceness of word and reference image (first image) is given by the MAE value. The small errors signify that the image is close to the reference image. An average of MAE is taken from the value of overall results.

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

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.

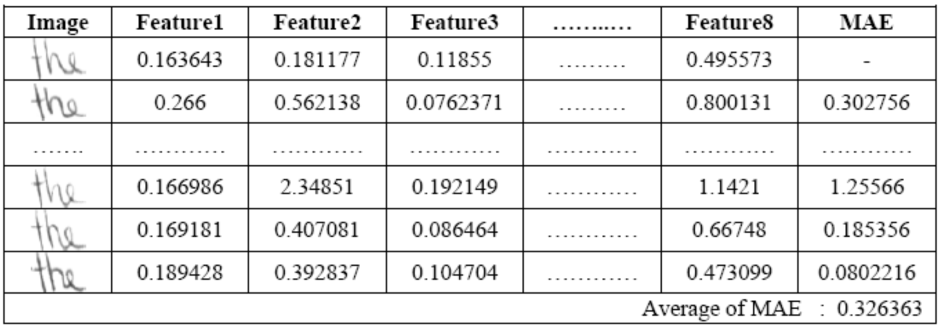


Image	Feature1	Feature2	Feature3	………..…	Feature8	MAE
the	0.163643	0.181177	0.11855	………	0.495573	-
the	0.266	0.562138	0.0762371	………	0.800131	0.302756
…….	…………	…………	…………	…………	…………	…………
the	0.166986	2.34851	0.192149	…………	1.1421	1.25566
the	0.169181	0.407081	0.086464	…………	0.66748	0.185356
the	0.189428	0.392837	0.104704	…………	0.473099	0.0802216
					Average of MAE :	0.326363

Tab. I *Example of MAE calculation.*

Authorship invarianceness for the invariant feature vector and discretized feature vector is calculated by performing the intra-class and inter-class analysis of MAE value. The analysis result shows that the variance between feature for intra-class (same writer) and inter-class (different writer) using the discretized feature vector gives a better result compared to undiscretized data. It has improved the authorship invarianceness where the MAE value for intra-class using discretized data is smaller than undiscretized data, and MAE value for inter-class using discretized data is higher than undiscretized data. The lowest MAE value in intra-class indicates the features are most similar to each other for the same writer whilst the highest MAE value for intra-class indicates they are most different to each other for different writers. These results have proved the hypothesis that the proposed invariant discretization can improve the authorship invarianceness with a standard representation of individual features for the individuality representation. Fig. 7 and Fig. 8 below show the comparison of authorship invarianceness for the UMI technique with discretized data and undiscretized data for various words and a similar word, respectively.

Uniqueness of individual features for each writer is described by the result in Fig. 7 and Fig. 8. It satisfies the individuality of handwriting concept where the MAE value for intra-class (same writer) is lower than inter-class (different writers). More important is that the individual features are better illustrated by using

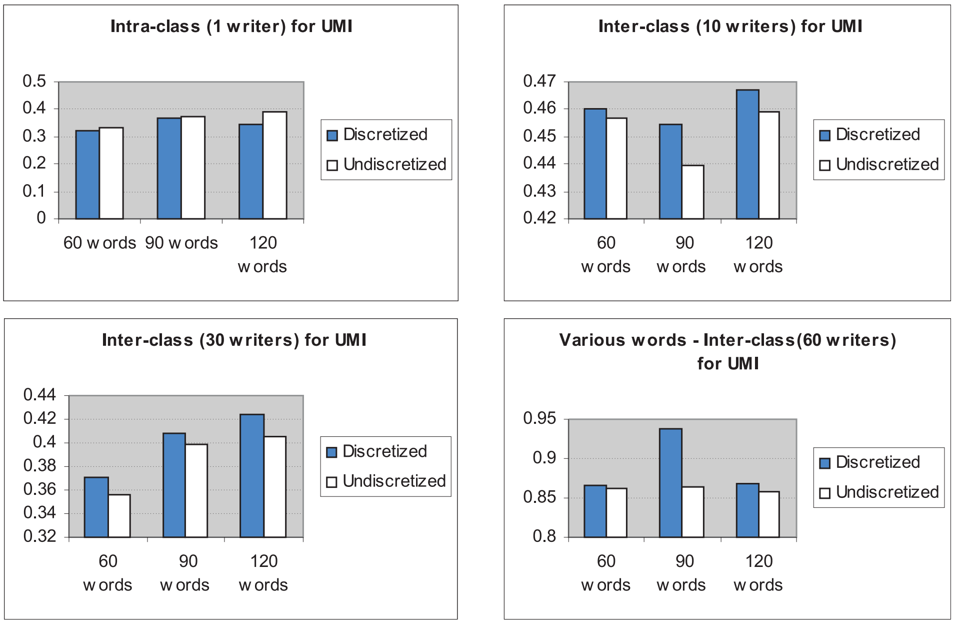


Fig. 7 *Authorship invarianceness comparison for various words.*

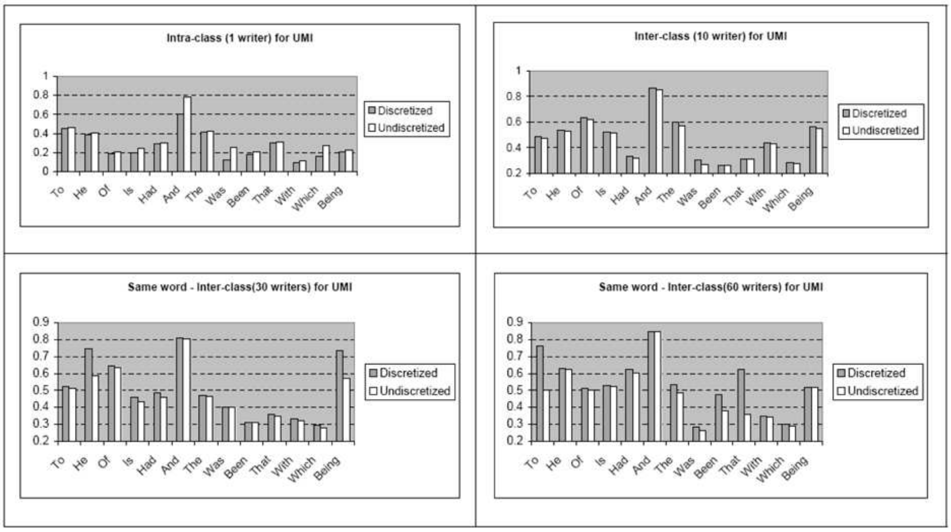


Fig. 8 *Authorship invarianceness comparison for a similar word.*

the discretized feature vector when compared to the undiscretized feature vector. The discretized data should have a lower MAE value than the undiscretized data for intra-class (same writer), and the discretized data should give a higher MAE value when compared to the undiscretized data for inter-class (different writers). Furthermore, results for a large number of data are also provided in Fig. 9 to show the discretized data has improved the MAE value for inter-class. However, the intra-class experiment cannot be performed because only limited data can be prepared for a single writer. All illustrations in Figs. 7–9 satisfy the individuality of handwriting concept and show improvement. Thus, it is proven that the proposed invariant discretization gives better authorship invarianceness when compared to the undiscretized data.

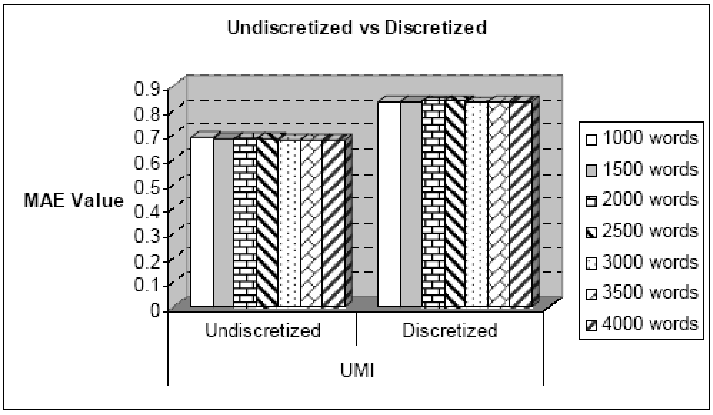


Fig. 9 *Undiscretized vs discretized data for inter-class.*

6.2 Identification performance

An experiment has been conducted to evaluate identification performance using the proposed invariant discretization and different discretization techniques in the Rosetta (Rough Set Toolkit) [45]. The comparisons are done with undiscretized data. 4400 data have been divided into 5 data sets in order to form training and testing data sets for the identification task, as shown in Fig. 10.

From the above figure, two data sets –SET 12345 and SET 13524 – have been prepared. Each set consists of three more data sets; (i) 3520 training data with 880 testing data, (ii) 2640 training data with 1760 testing data, and (iii) 2200 training data with 2200 testing data. Three discretization techniques in the Rosetta Toolkit are implemented to obtain accuracy as shown in Table 2 (SET 12345) and Table 3 (SET 13524). These include Naive (Naive Algorithm), Semi-Naive (Semi-Naive Algorithm) and Boolean (Boolean Reasoning Algorithm). On the other hand, InvDis is a label for Invariant Discretization and UnDis means Undiscretized Data. In the Rosetta Toolkit, we used GA (Genetic Algorithm), John (Johnson's Algorithm) and 1R (Holte's 1R Algorithm) as rules reduction prior to the classification.

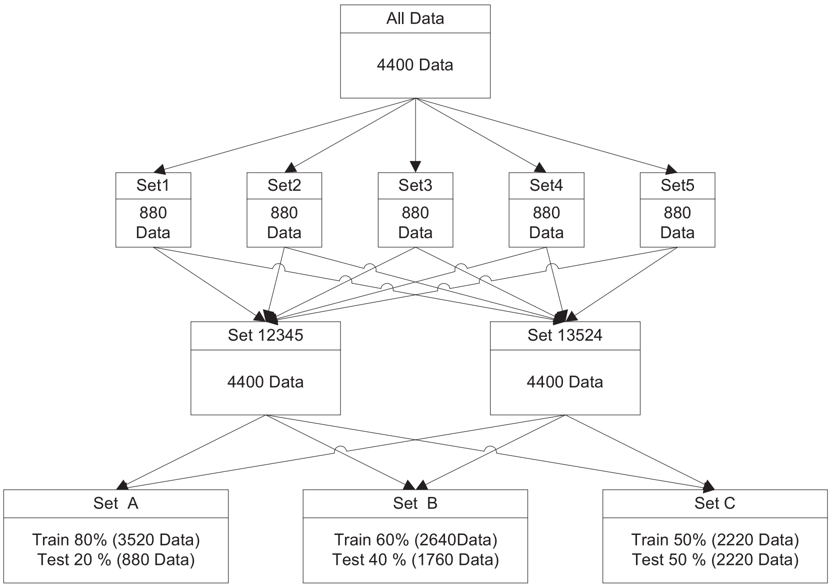


Fig. 10 *Data collection for training and testing.*

SET 12345	**Reduction Discretize**	**GA**	**John**	**1R**
SET 1 3520 -Train (80%) 880 - Test (20%)	Naive	99.97	99.89	99.97
	Semi-naive	99.97	99.89	99.97
	Boolean	99.20	99.20	20.48
	UnDis	33.56	33.56	33.67
	InvDis	99.97	99.09	99.97
SET 2 2640 -Train (60%) 1760 - Test (40%)	Naive	99.49	99.32	99.49
	Semi-naive	99.49	99.15	99.43
	Boolean	98.58	98.58	14.68
	UnDis	30.55	30.55	30.66
	InvDis	99.97	98.75	99.97
SET 3 2200 -Train (50%) 2200 - Test (50%)	Naive	99.0	98.82	99.0
	Semi-naive	99.0	98.86	98.91
	Boolean	97.64	97.64	14.45
	UnDis	29.49	29.49	29.53
	InvDis	99.97	98.55	99.97

Tab. II *Comparison of accuracy for various discretization techniques using data of SET 12345.*

SET 13524	**Reduction Discretize**	**GA**	**John**	**1R**
SET 1 3520 -Train (80%) 880 - Test (20%)	Naive	99.77	97.61	99.77
	Semi-naive	99.77	98.64	99.77
	Boolean	97.05	97.05	21.02
	UnDis	34.62	34.62	34.73
	InvDis	**99.95**	**99.56**	**99.95**
SET 2 2640 -Train (60%) 1760 - Test (40%)	Naive	99.89	99.32	99.89
	Semi-naive	99.89	98.69	98.69
	Boolean	97.44	97.44	18.41
	UnDis	29.92	29.92	30.03
	InvDis	**99.95**	**97.95**	**99.95**
SET 3 2200 -Train (50%) 2200 - Test (50%)	Naive	98.18	98.04	98.18
	Semi-naive	98.18	98.09	98.18
	Boolean	96.77	96.77	14.42
	UnDis	26.78	26.78	26.88
	InvDis	**99.95**	**98.18**	**99.95**

Tab. III *Comparison of accuracy for various discretization techniques using data of SET 13524.*

Both tables show the accuracy of discretized data is higher when compared to undiscretized data, except for Boolean discretization with IR reduction. This is due to the variance between features that have been improved by implementing the discretization technique. Authorship invarianceness is improved where the variance between features for intra-class using discretized features is lower than undiscretized data. Meanwhile, the variance between features for inter-class using discretized features is higher than undiscretized data. Individual features are clustered into the same interval that explicitly corresponds to the same writer, which directly contributed to better identification performance.

7. Conclusion

The results of MAE value in Section 6.1 suggest that the invarianceness of authorship using the discretized feature vector for intra-class (same writer) and inter-class (different writers) is improved compared to the undiscretized feature vector. It satisfies the individuality of handwriting concept in WI, where the MAE value for intra-class (same writer) should have a smaller value and the MAE value for inter-class (different writers) should be higher, regardless of any types of words tested. The discretized feature vector obtained from the proposed invariant discretization process has also shown that each writer has its own representations illustrated by the discretized feature vector. It represents the individuality of handwriting concept in the WI domain where each person has its own style of writing. The standard representation for each writer makes smaller variance between features for intra-class (same writer) and larger for inter-class (different writers), compared to the original extracted invariant feature vectors. This directly contributed to better performance for individuality of handwriting, as provided in Section 6.2. Thus, this authorship invarianceness analysis confirms that the proposed invariant discretization is worth further exploration in the WI domain.

References

[1] Srihari S. N., Huang C., Srinivasan H., Shah V. A.: Biometric and forensic aspects of digital document processing, Digital Document Processing, B. B. Chaudhuri (ed.), Springer, 2006.

[2] Tapiador M., Sigüenza J. A.: Writer Identification Method Based on Forensic Knowledge. First International Conference on Biometric Authentication, ICBA 2004, 2004, pp. 555-561.

[3] Kun Y., Yunhong W., Tieniu T.: Writer identification using dynamic features, Biometric Authentication: First International Conference, ICBA 2004, Hong Kong, China, July 15-17, 2004, pp. 512-518.

[4] Yong Z., Tieniu T., Yunhong W.: Biometric personal identification based on handwriting, Pattern Recognition, Proc. 15th International Conference on Volume 2, 3-7 Sept 2000, pp. 797-800.

[5] Somaya M., Eman M., Dori K., Fatma M.: Writer Identification Using Edge-based Directional Probability Distribution Features for Arabic Words, IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2008, pp. 582-590.

[6] Niels R., Vuurpijl L., Schomaker L.: Automatic allograph matching in forensic writer identification, International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI). **21**, 1, 2007, pp. 61-81.

[7] Pervouchine V., Leedham G., Melikhov K.: Handwritten Character Skeletonisation for Forensic Document Analysis, Proceedings of the 2005 ACM symposium on Applied computing, Santa Fe, New Mexico, USA.

[8] Franke K., Koppen M.: A Computer-based System to Support Forensic Studies on Handwritten Documents, International Journal on Document Analysis and Recognition, **3**, 4, 2001, pp. 218-231.

[9] He Z.-Y, You X., Tang Y.-Y.: Writer Identification Using Global Wavelet-based Features, Neurocomputing, **71**, 10-12, 2008, pp. 1832-1841.

[10] Zhang, B., Srihari S. N.: Analysis of Handwriting Individuality Using Word Features, Proceedings of the Seventh International Conference of Document Analysis and Recognition, 2003, pp. 1142-1146.

[11] Srihari S. N., Cha S.-H., Arora H., Lee S.: Individuality of Handwriting, Journal of Forensic Sciences, **47**, 4, 2002, pp. 1-17.

[12] Srihari S. N., Cha S.-H., Lee S.: Establishing Handwriting Individuality Using Pattern Recognition Techniques, Proceedings of Sixth International Conference on Document Analysis and Recognition, 2001, pp. 1195-1204.

[13] Srihari S. N., Cha S.-H., Arora H., Lee, S.: Individuality of Handwriting: A validation study, Sixth IAPR International Conference on Document Analysis and Recognition, Seattle, 2001.

[14] Yong Z., Tieniu T., Yunhong W.: Biometric Personal Identification Based on Handwriting, Proceeding of 15th International Conference on Pattern Recognition, **2**, 2000, pp. 797-800.

[15] Leedham G., Chachra S.: Writer Identification using Innovative Binarised Features of Handwritten Numerals, Proceeding of Seventh International Conference of Document Analysis and Recognition, **1**, 2003, pp. 413-416.

[16] Zois E. N., Anastassopoulos V.: Morphological Waveform Coding for Writer Identification, Pattern Recognition, **33**, 3, 2000, pp. 385-398.

[17] Muda A. K., Shamsuddin S. M., Darus M.: Invariants Discretization for Individuality Representation in Handwritten Authorship, International Workshop on Computational Forensic (IWCF 2008), LNCS 5158, Springer Verlag, pp. 218-228.

[18] Agre G., Peev S.: On Supervised and Unsupervised Discretization, Cybernetics and Information Technologies, **2**, 2, 2002, pp. 43-57.

[19] Nguyen H. S.: Discretization Problems for Rough Set Methods, Rough Sets & Current Trend in Computing, First International Conference of RSCTC'98, LNAI 1424, Warsaw, Poland, 1998, pp. 545-552.

[20] Xin G., Xiao Y., You H:, Discretization of Continuous Interval-Valued Attributes in Rough Set Theory and Application, International Conference on Machine Learning and Cybernetics, **7**, 2007, pp. 3682-3686.

[21] Liu H., Hussain F., Tan C.-L., Dash M.: Discretization: An Enabling Technique, Data mining and Knowledge Discovery, **6**, 2002, pp. 393-423.

[22] Dougherty J., Kohavi R., Sahami M.: Supervised and Unsupervised Discretization of Continuous Features. Twelfth International Conference on Machine Learning. Los Altos, CA: Morgan Kaufmann, 1995, pp. 194-202.

[23] Chmielewski M. R., Grzymala-Busse J. W.: Global Discretization of Continuous Attributes as Preprocessing for Machine Learning, International Journal of Approximate Reasoning, **15**, 4, 1996, pp. 319-331.

[24] Liao S. X: Image Analysis by Moment. University of Manitoba: Ph.D. Dissertation, 1993.

[25] Hu M.-K.: Visual Pattern Recognition by Moment Invariants, IRE Transaction on Information Theory. **8**, 2, 1962, pp. 179-187.

[26] Alt F. L.: Digital Pattern Recognition by Moments, Journal of the ACM (JACM), **9**, 2, 1962, pp. 240-258.

[27] Yinan S., Weijun L., Yuechao W.: United Moment Invariant for Shape Discrimantion, IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, 2003, pp. 88-93.

[28] Chen C.-C.: Improved moment invariants for shape discrimination, Pattern Recognition, **26**, 5, May 1993, pp. 683-686.

[29] Liu C.-L., Nakashima K., Sako H., Fujisawa H.: Handwritten Digit Recognition: Benchmarking of State-of-the-art Techniques, Pattern Recognition, **36**, 10, 2003, pp. 2271-2285.

[30] Bulacu M., Schomaker L., Brink A.: Text-independent Writer Identification and Verification on Offline Arabic Handwriting, Proc. of 9th Int. Conf. on Document Analysis and Recognition (ICDAR 2007), IEEE Computer Society, Curitiba, Brazil. **2**, 2007, pp. 769-773.

[31] Bulacu M., Schomaker L.: Combining Multiple Features for Text-Independent Writer Identification and Verification, Proc. of 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR 2006), La Baule, France, 2006, pp. 281-286.

[32] Schlapbach A., Bunke H.: Off-line Handwriting Identification Using Gaussian Mixture Models. In: Proc. 18th Int. Conf. on Pattern Recognition (ICPR06), 2006, pp. 992–995.

[33] Tomai C. I., Zhang B., Srihari S. N.: Discriminatory Power of Handwritten Words for Writer Identification, Proc. Int. Conf. on Pattern Recognition (ICPR 2004), Cambridge, England, 2004.

[34] Dehkordi M. E., Sherkat N., Allen T.: Handwriting Style Classification, International Journal on Document Analysis and Recognition, **6**, 1, 2003, pp. 55-74.

[35] Said H. E. S., Tan T. N., Baker K. D.: Writer Identification Based on Handwriting, Pattern Recognition, **33**, 2000, pp. 149-160.

[36] Liu C.-L., Dai R.-W, Liu Y.-J.: Extracting Individual Features from Moments for Chinese Writer Identification", Proceedings of the Third International Conference of Document Analysis and Recognition, **1**, 1995, pp. 438-441.

[37] Bulacu M., Schomaker L.: A Comparison of Clustering Methods for Writer Identification and Verification, Proc. of 8th Int. Conf. on Document Analysis and Recognition (ICDAR 2005), IEEE Computer Society, Seoul, Korea, **2**, 2005, pp. 1275-1279.

[38] Gunter S., Bunke H.: Ensembles of Classifiers for Handwritten Word Recognition Specialized on Individual Handwriting Styles, Proc. 6th Int. Workshop of Document Analysis Systems IV, Springer LNCS 3163, 2004, pp. 286-297.

[39] Cha Sung-Hyuk, Srihari S. N.: Writer Identification: Statistical Analysis and Dichotomizer, Advances in Pattern Recognition: Joint IAPR International Workshops, SSPR 2000 and SPR 2000, Alicante, Spain, 2000.

[40] Pervouchine V., Leedham G.: Extraction and Analysis of Forensic Document Examiner Features Used for Writer Identification, Pattern Recognition, **40**, 3, 2007, pp. 1004-1013.

[41] Brink A., Schomaker L., Bulacu M.: Towards Explainable Writer Verification and Identification Using Vantage Writers, Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR2007), Curitiba, Brazil, 2007, pp. 824-828.

[42] He Z.-Y., Tang Y.-Y, You X.: A Contourlet-based Method for Writer Identification, IEEE International Conference of Systems, Man and Cybernetics, **1**, 2005, pp. 364-368.

[43] Pervouchine V., Leedham G.: Extraction and Analysis of Document Features from Vector Skeletons of Grapheme 'th', DAS 2006, LNCS 3872, 2006, pp. 196-207.

[44] Gazzah S., Amara N. E. B.: Writer Identification Using Modular MLP Classifier and Genetic Algorithm for Optimal Features Selection, ISNN 2006, LNCS 3972, 2006, pp. 271-276.

[45] Ohrn A., Komorowski J.: ROSETTA: A Rough Set Toolkit for Analysis of Data, Proceeding of Third International Joint Conference on Information Sciences, Durham, **3**, 1997, pp. 403-407.

[46] Marti U.-V., Bunke H.: The IAM-database: an English Sentence Database for Off-line Handwriting Recognition, Int. Journal on Document Analysis and Recognition, **5**, 2002, pp. 39-46.