2009 International Conference on Biometrics and Kansei Engineering

Authorship Invarianceness for Writer Identification

Azah Kamilah Muda[1], Siti Mariyam Shamsuddin[2], Ajith Abraham[3]
Universiti Teknikal Malaysia Melaka, 75450 Melaka, Malaysia[1]
Universiti Teknologi Malaysia, 81310 Johor, Malaysia[2]
Machine Intelligence Research Labs (MIR Labs), USA[3]
azah@utem.edu.my; mariyam@utm.my; ajith.abraham@ieee.org

***Abstract*—The uniqueness of shape and style of handwriting can be used for author's authentication. Acquiring individual features to obtain Authorship Invarianceness Concept have led to an important research in Writer Identification domain. This paper discusses the investigation of this concept by extracting individual features using Geometric Moment Function. Experiment results have shown that Handwriting Invarianceness are discerning with better identification accuracy. This has verified that Moment Function is worth to be explored in identifying the handwritten authorship for Writer Identification.**

Keywords-Writer Identification; Authorship Invarianceness; Moment Function;

1.Introduction

Writer Identification (WI) 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]. It has a great importance in the criminal justice system and widely explored in forensic handwriting analysis [1],[5-8]. WI distinguishes writers based on the handwriting while ignoring the meaning of the word or character written.

The main issue in WI is how to acquire the features that reflect the author of handwriting. Many approaches have been proposed to extract the rigid characteristics of the shape such as in [2],[3],[9-14]. However, rigid characteristic contributed to the large lexicon. A common behavior of actual systems is that the accuracy decreases as the number of reference vector in the lexicon grows [15]. The computational complexity is also related to the lexicon, and it increases relatively to its size [16]. Meanwhile, the global approach does not incur additional lexicon into the database [17-19].

Rigid characteristics also lead to the various representations of a writer in handwriting and contributed to the large variation between features for intra-class and low variation for inter-class. Intra-class and inter-class variation are important in classification. Thus, this study focuses on extracting global features of word shape using Moment Function (MF) in order to represent the individual features of a writer.

This paper discussed the exploration of Individuality of Handwriting by extracting global features from handwritten word shape. Individual features are verified with the proposed Authorship Invarianceness of Moment Function in Writer identification. The remainder of the paper is structured as follows. In Section 2, an overview of handwriting individuality is discussed. Authorship Invarianceness method is described in Section 3, followed by the experiments in Section 4. Finally, conclusion and future work is drawn in Section 5.

2.Individuality of Handwriting

Handwriting is individualistic. It rests on the hypothesis that each individual has consistent handwriting [1],[4],[20-23]. The relation of character, shape and the style of writing are different from one to another. These various styles are required to classify in order to identify which group or classes that they are closed to. It must be unique feature that can be generalized as individual features. Figure 1 shows that each person has its individuality styles of writing. The shape is slightly different for the same writer and quite difference for different writers.

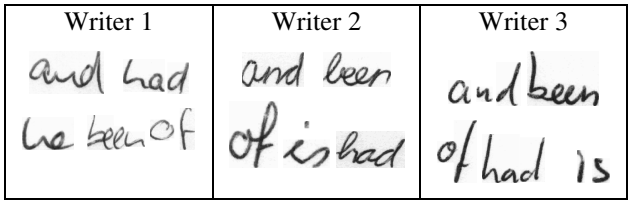

Figure 1. Different word for different writer.

3.Authorship Invarianceness

The concept of Authorship Invarianceness is proposed to validate the Handwriting Inviduality from the extracted features of MF in WI domain. This is important since acquiring the individual features is the main element in identification the handwritten authorship. Many previous works have developed new approaches or techniques to extract these individual features in WI domain. Therefore, the significant of the proposed method with United Representation technique to extract individual global features is important prior to its in-depth usage in WI.

The invarianceness in MF is defined as *preservation of the images regardless of its transformations* where it gives small similarity error for intra-class (same image) and large similarity error for inter-class (different image). This concept can be adopted into WI domain as *preservation of individual features for a writer regardless of his writings* to validate Handwriting Individuality prior to classification task. The

978-0-7695-3692-7/09 $25.00 © 2009 IEEE
DOI 10.1109/ICBAKE.2009.13

Authorship Invarianceness method is illustrated in the next section.

A. Authorship Invarianceness Procedure

Authorship Invarianceness Procedure consists of three processes: extracting global features from moment representation, similarity measurement of the variance between features and intra-class and inter-class analysis. These procedures are proposed to employ the MF and WI domain. MF is used to extract the individual features from global representation technique. While the concept of Handwriting Individuality is adopted into the intra-inter class analysis by calculating the similarity measurement; the variance between features (similarity error). Figure 2 illustrates the proposed procedure of Authorship Invarianceness.

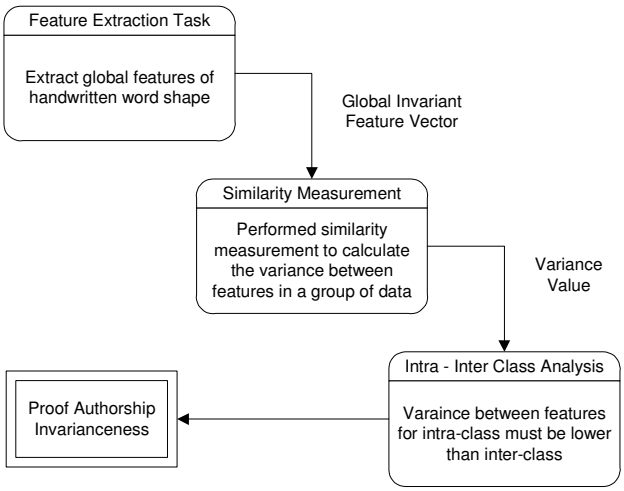

Figure 2. Authorship Invarianceness Procedure

B. Feature Extraction

MF has been used in diverse fields ranging from mechanics and statistics to pattern recognition and image understanding [24] for features extraction. Extensive usage of moments in image analysis and pattern recognition was inspired by Hu [25] and Alt [26]. MF is used for extracting global shape images. Shape is an important visual feature and it is one of the basic features for describing the image contents. However, to extract the features that represent and describe the shape precisely is a difficult task.

A good shape descriptor should be able to find perceptually similar shape that undergoes basic transformation, i.e., rotated, translated, scaled and affined transformed shapes. Due to the weaknesses of Hu's invariants [27], [28] proposed United Moment Invariant (UMI) where the rotation, translation and scaling can be discretely kept invariant to *region, closed and unclosed boundary*. The UMI provides a good set of discriminate shape features and valid in discrete condition. UMI is related to geometrical representation of Geometric Moment Invariant (GMI) [25] that considers normalized central moments as:

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

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

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

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

An Improved Moment Invariant (IMI) by Chen [29] is given as:

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

Equation (1), Equation (2) and Equation (3) have the factor μ_{pq}. By ignoring the influence of μ_{00} and ρ, UMI [28] is given as:

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

where ϕ_i are Hu's moment invariants.

C. Intra-Class vs Inter-Class

Individuality of Handwriting concept is proven with lower variance between features (similarity error) for intra-class (same writer) and higher in inter-class (different writer) class [22], [30-32]. This is due to the uniqueness of the extracted features in handwriting that called as individual features. As mentioned in [21], tow issues need to be addressed while comparing the handwriting: the variability of the handwriting of the same individual and the variability of the handwriting from one individual to another.

The within-writer variation is defined as the variation within a person's handwriting samples is less than the between-writer variation (the variation between the

handwriting samples of two different people). Mean Absolute Error (MAE) is performed in this work as the similarity measurer in Authorship Invarianceness to find the mean of variance between features in a group of data as shown below:

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

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's column of image.

MAE is used since it is corresponded to the Individuality of Handwriting measurement in WI domain. Each person will have specific features or characteristic in handwriting. By using the MAE, the variance between handwriting can be measured with similarity error of two handwritings from detail characteristics in feature's column. Smallest MAE value is considered as the most similar to original image which is the reference image to be compared. On the contrary, the highest MAE value is the most different. Therefore, the range of MAE between intra-class and inter-class is not a concern. As long as it proofs the characteristics of Handwriting Individuality Concept (the intra-class value must be lower than inter-class value).

4.Experimental Results

Two types of experiments have been conducted in this paper. First experiment is to validate MF can be used to extract individual features by using the handwriting invarianceness. MF of UMI has been explored in this experiment. The other one is to evaluate the performance of identification in classification task by using Rosseta Toolkit [34]. The experiments conducted in this paper used IAM database [33] with 4400 various images from 60 writers.

A. Handwriting Invarianceness

This section presents the result of Handwriting Invarianceness using UMI technique. The similarity measurement is calculated by using MAE (Equation 1). Example of MAE calculation is presented in Table I. The number of images is 20 for one author. Feature 1 to Feature 8 are extracted invariants that representing each word. The invarianceness of each word can be interpreted from the given MAE values with the same reference image (first image). The small errors signify that the image is close to the original image. An average of MAE is taken as the value of overall results.

Table 1. United Moment Invariant for 'the'

Image	Feature1	…….	Feature8	MAE
	0.163643	……	0.495573	-
	0.266	……	0.800131	0.302756
…….	……	……	……	……
	0.166986	……	1.1421	1.25566
	0.169181	……	0.66748	0.185356
	0.189428	……	0.473099	0.0802216
			Average of MAE :	0.326363

Tables 2 and 3 illustrate the MAE value for UMI using various words and same words, respectively. The variation of shape and style of writing for one writer (intra-class) is smaller compared to different writers (inter-class). This indicates the invarianceness between features for the same writer is smaller compared to different writers. Thus, it conforms that the UMI can be applied in WI domain.

Table 2. Invarianceness of Authorship for Various Words

Various Words	Intra-class 1 writer	Inter-class 10 writers	Inter-class 20 writers	Inter-class 60 writers
60	0.331508	0.456814	0.356398	0.861128
90	0.375977	0.439712	0.398261	0.864583
120	0.393223	0.459071	0.404881	0.857114

Table 3. Invarianceness of Authorship for Same Words

Word	Intra-class 1 writer	Inter-class 10 writers	Inter-class 20 writers	Inter-class 60 writers
To	0.465232	0.472446	0.511082	0.502622
He	0.405854	0.531652	0.58899	0.626304
Of	0.21199	0.617817	0.632907	0.502972
Is	0.244491	0.517127	0.433202	0.525358
Had	0.304395	0.317685	0.461475	0.601659
And	0.78404	0.851186	0.802919	0.847845
The	0.421582	0.573675	0.466353	0.483517
Was	0.25831	0.264641	0.401021	0.263911
Been	0.207651	0.262992	0.308957	0.380296
That	0.309031	0.311301	0.348014	0.35865
With	0.115092	0.428501	0.317997	0.341126
Which	0.269169	0.274032	0.280052	0.289756
Being	0.230172	0.550473	0.571301	0.51579

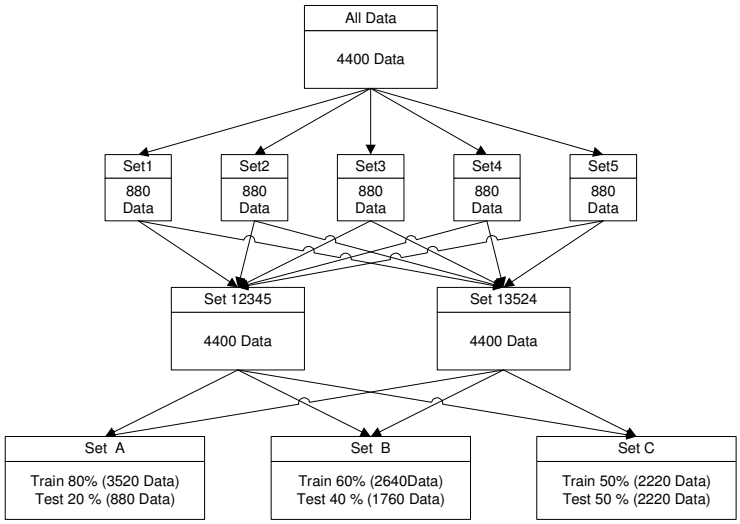

Figure 3. Data Collection for Training and Testing

B. Classification Accuracy

The experiment has been conducted to evaluate identification performance by implementing different discretization techniques using Rosetta (Rough Set Toolkit) [34] and proposed Invariant Discretization [35]. The effectiveness of employing discretization in improving identification performance of handwriting authorship is discussed in detailed in [35]. This paper only focuses on the evaluation of the extraction of individual features for Authorship Invarianceness. The comparisons are done with non-discretized data.

4400 data have been divided into 5 sets of data in order to form the training and testing data set for identification task, as shown in Figure 3. Two sets of data which are SET 12345 and SET 13524 have been prepared. Each of it consists of three datasets; (i) 3520 training data with 880 testing data (ii) 2640 training data with 1760 testing data (iii) 2200 training data with 2200 testing data. Three discretization techniques in Rosetta Toolkit are implemented to obtain the accuracy in Table 4 (SET 12345) and Table 5 (SET 13524). These include Naïve (Naïve Algorithm), Semi-Naïve (Semi_Naive Algorithm) and Boolean (Boolean Reasoning Algorithm). On the other hand, InvDis is the label for Invariant Discretization proposed by [35] and UnDis is meant as non-discretize data. In Rosetta Toolkit, we used GA (Genetic Algorithm), John (Johnson's Algorithm) and 1R (Holte's 1R Algorithm) as rules reduction prior to classification.

Table 4. Comparison of Accuracy for Various Discretization Technique Using Data of SET 12345

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

Table 5. Comparison of Accuracy for Various Discretization Technique Using Data of SET 13524

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**

Both Tables 4 and 5 show the accuracy of discretized data are higher compared to non-discretized data except Boolean discretization with IR reduction. This is due to the variance between features that have been improved by implementing discretization technique subsequent to feature extraction with MF. These features are clustered into the same cut that explicitly corresponds to the same author. The lower variation of intra-class and higher inter-class contributed to the better identification performance

5.Conclusions

This paper proposed Authorship Invarianceness method in order to validate MF in extracting individual features for WI domain. The experiments of UMI are performed to validate the handwriting invarianceness, and the extracted features are discretized for better identification. The results confirm that the invarianceness of handwriting is still preserved.

Acknowledgments

This work is supported by Ministry of Higher Education (MOHE) under Fundamental Research Grant Scheme (FRGS VOT 78182). Authors would like to thank Research Management Centre (RMC) Universiti Teknologi Malaysia, for the research activities and *Soft Computing Research Group* (SCRG) for the support in making this study a success.

References

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

[2] M. Tapiador, J.A. Sigüenza, "Writer identification method based on forensic knowledge," Biometric Authentication: First International Conference, ICBA 2004, Hong Kong, China, July 2004.

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

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

[5] M. Somaya, M. Eman, K. Dori, and M. Fatma, "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] R. Niels, L. Vuurpijl and L. Schomaker, "Automatic allograph matching in forensic writer identification," International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI). Vol. 21. No. 1.2007, pp. 61-81.

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

[8] K. Franke and M. Koppen, "A Computer-based System to Support Forensic Studies on Handwritten Documents," International Journal on Document Analysis and Recognition, Vol. 3. No. 4. 2001, pp. 218 – 231.

[9] A. Bensefia, T. Paquet and L. Heutte, "A Writer Identification and Verification System," Pattern Recognition Letters, In Press, Corrected Proof, Available online 23 May 2005.

[10] A. Schlapbach and H. Bunke, "Using HMM Based Recognizers for Writer Identification and Verification," In Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition, 2004, pp. 167–172.

[11] Z.-Y. He and Y.-Y. Tang, "Chinese Handwriting-based Writer Identification by Texture Analysis," Proceedings of International Conference of Machine Learning and Cybernetics, Vol. 6. 2004, pp. 3488 – 3491.

[12] C. Shen, X.-G. Ruan, and T.-L. Mao, "Writer Identification Using Gabor Wavelet," Proceedings of the 4th World Congress on Intelligent Control and Automation, Vol. 3. 2001, pp. 2061 – 2064.

[13] H.E.S Said, T.N. Tan, and K.D. Baker, "Writer Identification Based on Handwriting," Pattern Recognition, Vol. 33. 2002, pp.149-160.

[14] M. Wirotius, A. Seropian and N. Vincent, "Writer Identification from Gray Level Distribution," Proceeding of Seventh International Conference on Document Analysis and Recognition, 2002, pp. 1168 – 1172.

[15] A. L. Koerich, "Large Vocabulary Off-line Handwritten Word Recognition," Ecole de Technologie Superieure : Ph.D. Dissertation, 2002.

[16] A. L. Koerich, R. Sabourin and C.Y. Suen, "Large vocabulary off-line handwriting recognition: A survey," Pattern Anal Applic, Vol. 6. 2003, pp. 97–121.

[17] R.D. Cajote and R.C.L. Guevara, "Global Word Shape Processing Using Polar-radii Graphs for Offline Handwriting Recognition," IEEE Region 10 Conference of TENCON 2004. Vol. A. No. 1. 2004, pp. 315 – 318.

[18] M.E. Morita, "Automatic Recognition of Handwritten Dates on Brazilian Bank Cheque," Ecole de Technologie Superieure : Ph.D. Dissertation, 2003.

[19] S. Madvanath, and V. Govindaraju, "The Role of Holistic Paradigms in Handwritten Word Recognition," IEEE Transactions of Pattern Analysis and Machine Intelligence, Vol. 23. No.2. 2001, pp. 149-164.

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

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

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

[23] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee, "Individuality of Handwriting," Journal of Forensic Sciences, Vol. 47. No. 4.2002, pp. 1-17.

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

[25] M.-K. Hu, "Visual Pattern Recognition by Moment Invariants," IRE Transaction on Information Theory. Vol. 8. No. 2. 1962, pp. 179-187.

[26] F. L. Alt, "Digital Pattern Recognition by Moments," Journal of the ACM (JACM), Vol.9, No. 2. 1962, pp. 240 – 258.

[27] S.M. Shamsuddin, M.N. Sulaiman, and M. Darus, "Feature Extraction with an Improved Scale-Invariants for Deformations Digits," International Journal of Computer Mathematics, UK, Vol. 76, 2000, pp. 13-23, Taylor and Francis Group.

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

[29] C.-C. Chen, "Improved moment invariants for shape discrimination," Pattern Recognition, Volume 26, Issue 5, May 1993, pp. 683-686.

[30] Z. He, X. Youb and Y.-Y. Tang, "Writer Identification using Global Wavelet-based Features," Neurocomputing, Vol. 71. No. 10-12, 2008, pp. 1832-1841.

[31] G. Leedham and S. Chachra, "Writer Identification using Innovative Binarised Features of Handwritten Numerals," Proceeding of Seventh International Conference of Document Analysis and Recognition, Vol. 1.2003, pp. 413 – 416.

[32] E.N. Zois and V. Anastassopoulos, "Morphological Waveform Coding for Writer Identification," Pattern Recognition, Vol. 33. No.3. 2000, pp. 385-398.

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

[34] A. Ohrn and J. Komorowski, "ROSETTA: A Rough Set Toolkit for Analysis of Data," Proceeding of Third International Joint Conference on Information Sciences, Durham, Vol. 3. 1997, pp. 403–407.

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