Proceedings of the Fourth IASTED International Conference
Advances in Computer Science and Technology (ACST 2008)
April 2-4, 2008 Langkawi, Malaysia
ISBN CD: 978-0-88986-730-7

DISCRETIZATION OF INTEGRATED MOMENT INVARIANTS FOR WRITER IDENTIFICATION

Azah Kamilah Muda , Siti Mariyam Shamsuddin & Maslina Darus
Faculty of Computer Science & Information System,
University Technology of Malaysia, 81310 Skudai, Johor, Malaysia.
azah@utem.edu.my; mariyam@utm.my; maslina@pkrisc.cc.ukm.my

ABSTRACT
Conservative regular moments have been proven to exhibit some shortcomings in the original formulations of moment functions in terms of scaling factor. Hence, an incorporated scaling factor of geometric functions into United Moment Invariant function is proposed for mining the feature of unconstrained words. Subsequently, the discrete proposed features undertake discretization procedure prior to classification for better feature representation and splendid classification accuracy. Collectively, discrete values are finite intervals in a continuous spectrum of values and well known to play important roles in data mining and knowledge discovery. Many induction algorithms found in the literature requires that training data contains only discrete features and some works better on discretized data; in particular rule based approaches like rough sets. Hence, in this study, an integrated scaling formulation of Aspect Scaling Invariant is presented in Writer Identification to hunt for the individuality perseverance. Successive exploration is executed to investigate for the suitability of discretization techniques in probing the issues of writer authorship. Mathematical proving and results of computer simulations are embraced to attest the feasibility of the proposed technique in Writer Identification. The results disclose that the proposed discretized invariants reveal 99% accuracy of classification by using 3520 training data and 880 testing data.

KEY WORDS
Writer identification, scaling factor, discretization, moment function.

1. Introduction

Global Moment Function can be used to produce a set of moments that uniquely embody a global characteristic of image shape. It has been used in various fields ranging from mechanics and statistics to pattern recognition and image understanding [1]. The exploiting of moments in image analysis and pattern recognition was inspired by Hu [2], and Alt [3]. Hu [2], presented a set of seven-tuplet moments that invariant to position, size, and orientation of the image shape. Nevertheless, there are many research have been done to proof that there were some drawbacks in the original work by Hu [2] in terms of invariant such as Reiss [4], Belkasim [5], Feng [6], Sivaramakrishna [7], Palaniappan [8] and Shamsuddin [9].

They proposed their method of moment and tested on feature extraction phase to represent the image. A good shape descriptor should be able to find perceptually similar shape regardless of basic transformation; rotation, translation, scaling and affined transformed shapes. Furthermore, it can abide with human beings in comparing the image shapes. Yinan [10] derived United Moment Invariants (UMI) based on basic scaling transformation by Hu [2]. It can be applied in all conditions with a good set of discriminating shape features. With UMI competency as a good description of image shape, integrated scaling transformation of Aspect Invariant Moment (ASI) [6] with UMI [10] is developed for Writer Identification (WI). From our reviews, UMI has never been tested in WI domain.

WI can be embraced as a meticulous kind of dynamic biometric in pattern recognition for forensic application. The shapes and writing styles can be used as biometric features for identifying an identity Srihari [11], Tapiador [12], Kun [13], Yong [14]. WI discerns writers based on the writing styles and the shape of writing. There are two types of classes that should be considered while comparing handwriting which are intra-class (same writer) and inter-class (different writer). Similarity error for inter-class should be higher than intra-class for invariancess of authorship. These similarity errors can be associated to discretization technique by discerning the images in categorizing the individual features. The continuous values of invariant features are discretized to obtain the detachment of authors' individuality for better classification.

The remainder of this paper is systematized as follows. An overview of United Moment Invariant and Geometric Scaling of Aspect Invariant Scaling is given in Section 3, followed by integrated scaling factor of ASI and UMI.

Section 4 briefly depicts the rough sets theory and discretization process. Section 5 illustrates the implementation and the results of the proposed technique, and finally, conclusion and future work is given in Section 6.

2. United Moment Invariant

Yinan [10] proposed UMI with a good quality set of shape features and compelling in discrete condition. Additionally, rotation, translation and scaling can be distinctly kept invariant to region, closed and unclosed boundary using UMI. In terms of mathematics, UMI is related to Geometric Moment Invariants (GMI) by considering Equation (1) as normalized central moments in Hu [2]:

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

$$p+q = 2,3,\ldots. \qquad (1)$$

and Equation (2) in discrete form. Normalized central moments are given as:

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

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

and improved moment invariant by Chen [15] given as:

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

Equation (1) to Equation (3) reveal that the formulations have the factor μ_{pq}. By eliminating this factor, the eight feature vector can be derived and it is shown as below, where ϕ_i are Hu's moment invariants [10].

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

3. Geometric Scaling Invariant

Hu [2] presented moment invariants in 2-D pattern recognition from the first three central moments, utterly tested on automatic character recognition. He alleged that the generated set of moments are invariant to position, size, and orientation of the image shape. Nonetheless, his approach couldn't furnish for unconstrained scaling images ([5],[6],[7],[8],[9],[16], [17]).

3.1 Aspect Scaling Invariant (ASI)

In accordance to Feng [6], the proposed GMI proposed by Hu [2] has several shortcomings. Hence, Feng [6] proposed Aspect Scaling Invariant (ASI) for unequal scaling images by structuring moment invariants which are independent of the diverse scaling in the *x* and *y* directions. It is given as:

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

The numerator and denominator of the scaling factor are in same order. Consequently, the magnitude of the aspect invariant moments will not vary dramatically with moment order. This consented to the effectiveness of using high order moments for augmenting the discrimination ability of the system.

3.2 An Integrated of ASI into UMI

By considering Equation (2) in discrete form, UMI can be depicted to relate with GMI by Hu [2], which is the Normalized Central Moments in Equation (2) and improved moment invariant in Equation (3) by Chen [15]. As example, consider only $\theta_1 = \frac{\sqrt{\phi_2}}{\phi_1}$. From Hu, $\phi_2 = (\eta_{20}-\eta_{02})^2 + 4\eta_{11}^2$, substitute Normalized Central Moments (Equation (2)) into $\phi_2 = (\eta_{20}-\eta_{02})^2 + 4\eta_{11}^2$, we get:

$$\phi_2 = \left(\frac{\mu_{20}-\mu_{02}}{\mu_{00}^2}\right)^2 + \frac{4\mu_{11}^2}{\mu_{00}^4}. \qquad (6)$$

Substitute Equation (6) into Equation (2) yields,

$$\sqrt{\phi_2} = \frac{\sqrt{(\mu_{20}-\mu_{02})^2 + 4\mu_{11}^2}}{\mu_{00}^2}, \qquad (7)$$

and,

$$\phi_1 = \eta_{20} + \eta_{02} = \frac{\mu_{20} + \mu_{02}}{\mu_{00}^2}. \qquad (8)$$

Hence,

$$\frac{\sqrt{\phi_2}}{\phi_1} = \frac{\sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{\mu_{20} + \mu_{02}} = \theta_1. \qquad (9)$$

Equivalent process is assessed for different scaling factor, and in this study, we employ ASI to obtain the invarianceness of $\theta_1 = \theta_1' = \theta_1''$.

4. Discretization

Discretization of real value attributes is an essential task in data mining, predominantly the classification problem. Empirical results are showing that the superiority of classification methods depends on the discretization algorithm used in preprocessing step. In general, discretization is a process of exploring for partition of attribute domains into intervals and unifying the values over each interval. Discretization engages searching for "cuts' that determine intervals. All values that lie within each interval are mapped to the same value, in effect converting numerical attributes that can be treated as being symbolic Nguyen [18].

There are abundant of discretization algorithms exists, and based on different theoretical; probability statistics, Boolean reasoning, clustering and error/entropy calculations. Generally, discretization techniques can be divided into two classes:

- Unsupervised Discretizations – no classification information available for the objects being considered. These methods respond to the assumptions of the distribution of the attribute values.

- Supervised Discretizations – classification information is on hand, and this information can be taken into consideration when discretizing the data. A common denominator can be employed to minimize the number of objects from different decision classes into the same discretization class.

In this study, the effects of discretization on proposed integrated invariants feature for writer identification is investigated. Our results disclose that the performance of the classification on writers' handwriting is significantly improved when the proposed invariant features are discretized with unsupervised discretization algorithms.

5. Experiment Result

There are two types of experiments have been conducted in this paper. First experiment is to proof handwriting invarianceness where the proposed ASI into UMI can be used in handwriting individuality. The other one is to evaluate the performance of identification in classification task.

5.1 Handwriting Invarianceness

The proposed technique is analyzed on unconstrained handwritten words. The invarianceness of the proposed method is compared with the original GMI, UM and ASI into GMI using writer identification data. The issue in writer identification is to discover the individuality of handwriting for each writer that based on the adjacent unknown handwriting in the database. To accomplish this, we execute intra-class analysis to uncover the nearest words within the same class or the same writer with the lowest Mean Absolute Error (MAE) value. The MAE function is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(x_i - r_i)\right| \qquad (10)$$

Table 1 represents the invariants results of word '*the'* using UMI. The number of image 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 values of MAE with the same reference image (first image). The small errors signify that the image is closed to the original image. An average MAE is the value of overall result.

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

Table 1 : United Moment Invariant for 'the'

The MAE is computed for each moment technique, and the results are shown in Table 2 through Table 4.

Table 2 : MAE Comparison for 'the'

GMI	ASI -GMI
1.08545	0.69487
UMI	**ASI - UMI**
0.326363	0.708906

Table 3 : MAE Comparison for 'was'

GMI	ASI -GMI
0.676755	0.789657
UMI	**ASI - UMI**
0.261064	0.736664

Table 4 : MAE Comparison for 'that'

GMI	ASI -GMI
0.651049	0.74402
UMI	**ASI - UMI**
0.399031	0.677631

From the above tables, MAE for UMI furnishes the lowest mean value compared to the other moment techniques, thus it is chosen for representing the moment approaches. Even the proposed techniques are incomparable with the original UMI, the proposed techniques are able to corroborate the individuality concept in WI. Handwriting has long been considered individualistic and writer individuality rests on the hypothesis that each individual has consistent handwriting [11], [14], [19], [20], [21]. The variation of shape and style of writing in one same writer or intra-class are smaller compared to different writer or inter-class. **Table 5** and **Table 6** show the invarianceness of the same writer is smaller compared to different writer. Thus, it conforms to the proposed technique, and can be applied in WI domain.

Individuality of handwriting concept has been proof in many researches; Bin and Srihari [19], Srihari [21] and Liu [22]. However, the objective of this study is to create contributions towards leading scientific validation in WI using the proposed approaches. Additionally, UMI has never been tested in WI domain for the extraction of the WI individuality or authorship invarianceness. Hence, our proposed technique and UMI are worth for further exploration in WI domain.

Table 5 : Invariancess of Authorship using 'and'

Technique	Intra-class (1 writer)	Inter-class (10 writer)	Inter-class (20 writer)
Original UMI	0.351629	0.583265	0.405239
ASI Into UMI	0.893293	1.63293	1.55194

Table 6 : Invariancess of Authorship using 'that'

Technique	Intra-class (1 writer)	Inter-class (10 writer)	Inter-class (20 writer)
Original UMI	0.299813	0.457836	0.312279
ASI Into UMI	0.833782	1.30969	1.07363

5.2 Classification Accuracy

The experiment has been conducted to evaluate identification performance by implementing the different discretization techniques using rough set toolkit (Rosetta). The comparisons are done with undiscretized data. Three datasets have been used; (i) 3519 training data with 880 testing data (ii) 2640 training data with 1760 testing data (iii) 2200 training data with 2200 testing data. These data are extracted from 60 writers of AIM database [23]. **Table 7** illustrates the accuracy for each classification process. Naïve (Naïve Algorithm), Semi-Naïve (Semi_Naive Algorithm) and Boolean (Boolean Reasoning Algorithm) are discretization technique and UnDis is undiscretize data. Furthermore, GA (Genetic Algorithm), John (Johnson's Algorithm) and 1R (Holte's 1R Algorithm) are reduction of rules technique in Rosetta toolkit.

Table 7 : Comparison of Accuracy for Various Discretization Technique and UnDiscretized Data

Data	Reduction Discretize	GA	John	1R
SET 1 3520 -Train (80%) 880 - Test (20%)	Naive	99.97	99.32	99.97
	Semi-naive	99.97	99.09	99.97
	Boolean	99.43	99.43	32.39
	UnDis	35.34	35.34	35.34
SET 2 2640 -Train (60%) 1760 - Test (40%)	Naive	94.08	91.15	94.08
	Semi-naive	96.02	96.02	96.51
	Boolean	94.36	94.36	20.09
	UnDis	30.71	30.71	30.71
SET 3 2200 -Train (50%) 2200 - Test (50%)	Naive	93.58	93.86	93.58
	Semi-naive	95.39	95.39	95.75
	Boolean	93.93	93.93	20.06
	UnDis	27.59	27.59	27.59

Table 7 depicts that the accuracy of discretized data reveals higher accuracy compared to undiscretized data for the entire of discretization techniques excluding Boolean discretization with IR reduction. The higher accuracy in discretized data is due to the invarianceness of handwritten authorship as illustrated in the previous tables. The features are clustered in the same cut that explicitly corresponds to the same author. The lower variation of intra-class and higher inter-class contributed to this identification performance. Figure 1, Figure 2 and

Figure 3 give the comparisons of data SET 1, data SET2 and data SET3 respectively.

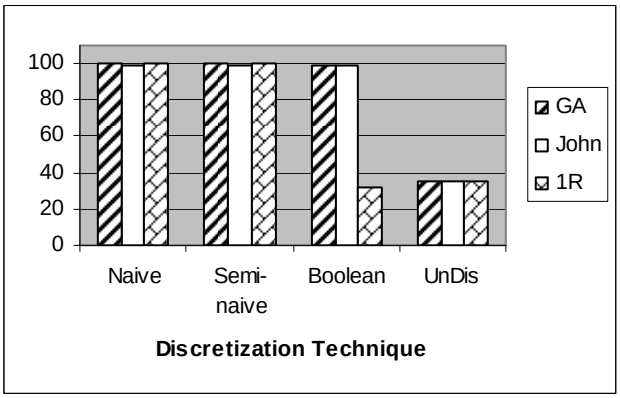

Figure 1: An accuracy for data SET 1

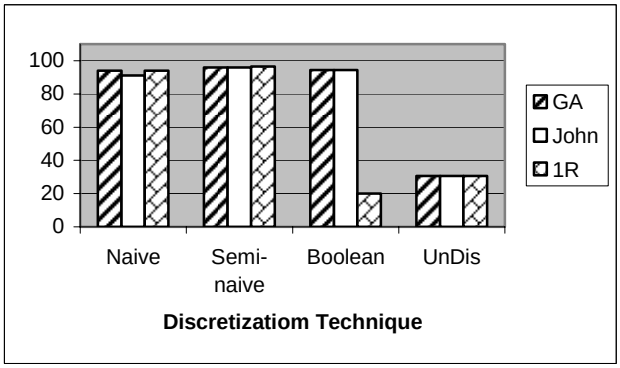

Figure 2: An accuracy for data SET 2

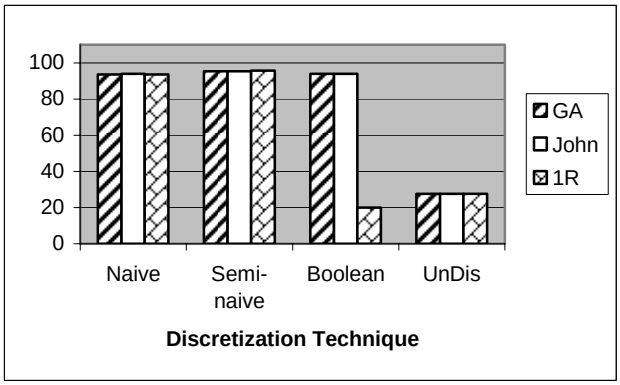

Figure 3: An accuracy for data SET 3

6. Conclusion and future work

This paper proposed an integrated scaling factor of Aspect Invariant Moment with United Moment Invariant for invarianceness of authorship in WI. The discrete features extracted from the proposed integrated invariants undergo discretization process for granular mining of writer authorship. Our experiments have revealed better results with various discretization techniques in identifying writer authorship. Despite higher MAE compared to UMI, the invarianceness is still conserved, thus conform to the theoretical concept of moment invariants. Comprehensive analysis on the effect of supervised and unsupervised discretization on the proposed integrated moment functions are still under probing for better identification and classification, particularly in the forensic document analysis and pattern recognition.

Acknowledgement

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

References

[1] X.L. Simon, Image analysis by moment, *Ph.D Thesis, University of Manitoba, Canada*, 1993.

[2] M.K. Hu, Visual pattern recognition by moment invariants, *IRE Transaction on Information Theory. 8*(2), 1962, 179 -187.

[3] F.L. Alt, Digital pattern recognition by moments, *Journal of the ACM (JACM), Volume 9*, Issue 2, April 1962, 240 - 258.

[4] T.H. Reiss, The revised fundamental theorem of moment invariants, *Pattern Analysis and Machine Intelligence, IEEE Transactions on Volume 13*, Issue 8, Aug. 1991 830 – 834.

[5] S.O. Belkasim, M. Shridhar, M. Ahmadi, Pattern recognition with moment invariants: a comparative study and new results, *Pattern Recognition, vol. 24*(12), 1991 1117-1138.

[6] P. Feng, M. Keane, A new set of moment invariants for handwritten numeral recognition, Image Processing, *Proc. ICIP-94., IEEE International Conference Volume 1*, 13-16 Nov. 1994, 154 -158.

[7] R. Sivaramakrishna, N.S. Shashidhar, Hu's moment invariant: how invariant are they under skew and perspective transformations?, *Conference on Communications, Power and Computing WESCANEX97 Proceedings*; Winnipeg, MB; May 22-23,1997, 292-295.

[8] R. Palaniappan, P. Raveendran, S. Omatu, New invariant moments for non-uniformly scaled images, *Pattern Analysis & Applications*, 3, 2000 78–87.

[9] S.M. Shamsuddin, M. Darus, M.N. Sulasian, Invarianceness of Higher Order Centralised Scaled-Invariants on Unconstrained Handwritten Digits, *International Journal of Inst. Maths. & Comp. Sciences (Comp. Sc. Ser), INDIA, 12*(1), 2001 1-9.

[10] S. Yinan, L. Weijun, W. Yuechao, United Moment Invariant for Shape Discrimination, IEEE *International Conference on Robotics, Intelligent Systems and Signal Processing*, China, Oktober, 2003 88-93.

[11] 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.

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

[13] 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, 512 – 518.

[14] 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, 797 – 800.

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

[16] P. Raveendran, S. Omatu, S.C. Poh, A new technique to derive invariant features for unequally scaled images, *Systems, Man, and Cybernetics, Computational Cybernetics and Simulation, 1997 IEEE International Conference on Volume 4*, 12-15 Oct. 1997 3158 – 3163.

[17] P. Raveendran, S. Omatu, A new technique to derive features for shift and unequally scaled images, Neural Networks, *Proc. IEEE International Conference on Volume 4*, 27 Nov -1 Dec. 1995, 2077 – 2080.

[18]H.S. Nguyen,Discretization problems for rough set methods, Rough Sets & Current Trend in Computing, *Lecture Notes in AI 1424, Proc. of the first International Conference, RSCTC'98*, Warsaw, Poland, June 1998, 545-552.

[19]B. Zhang, S.N. Srihari, Analysis of handwriting individuality using word features, Document Analysis and Recognition, *Proc. Seventh International Conference* on 3-6 Aug. 2003, 1142 - 1146.

[20] S.N. Srihari, S.-H. Cha, H. Arora, S. Lee, Individuality of handwriting, *Journal of Forensic Sciences, 47*(4), July 2002, 1-17.

[21] S.N. Srihari, S.-H. Cha, S. Lee, Establishing handwriting individuality using pattern recognition techniques, *Document Analysis and Recognition, Proc. Sixth International Conference on* 10-13 Sept. 2001, 1195 – 1204.

[22] C.-L. Liu, R.-W. Dai, Y.-J. Liu, Extracting individual features from moments for Chinese writer identification, *Document Analysis and Recognition, Proc. of the Third International Conference on Volume 1*, 14-16 Aug. 1995, 438 - 441 vol.1.

[23] U.-V Marti, H. Bunke, The IAM-database: an english sentence database for off-line handwriting recognition. *Int. Journal on Document Analysis and Recognition, Volume 5,* (2002) 39 – 46.