# Comparison of global and local features for author's identification by using geometrical and zoning methods

**I E A Jalil[1], S M Shamsuddin[2], A K Muda[1], M S Azmi[1], S Hasan[2] and S Ahmad[1]**

[1] Computational Intelligence and Technologies Lab, Centre for Advanced Computing Technology, Fakulti Teknologi Maklumat Dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), 76100, Melaka, Malaysia.
[2] Universiti Teknologi Malaysia (UTM), Skudai, 81300, Johor Bahru, Malaysia.

**Abstract.** Identification analysis for author's handwriting image in forensic investigation is still an important research area in this current big data era. Images feature extraction can lead to an issue of high dimensionality of data. The process of feature extraction is the most crucial process in author's identification. It is important to choose the best method to represent the image. This study compared two feature extraction methods, namely Higher-Order United Moment Invariant (HUMI) and the Edge-based Directional (ED) method that construct the Global and Local Features respectively. The additional process of discretization was implemented before the training and testing phase to represent the generalized features for the classifier models. This process induced a better performance accuracy for both methods where the discretized Local Features achieved 99.95% accuracy rate that slightly outperforms the discretized Global Features with only 99.91%.

## 1. Introduction

Research on extracting important and relevant features by any feature extraction method is still an important challenge in the area of machine learning [1, 2]. Most related machine learning researches especially in handling huge amount of data as such biometric data [3, 4, 5], prediction on stock exchange [6], prediction on software effort or fault [7, 8], traffic data [9] and image processing [10, 11, 26, 27] are to find the best technique to retrieve significant information to represent the data.

The process of extracting important features to represent the handwriting image starts from choosing the most suitable feature extraction methods. The production of features can be done in two types of features coverage either by local or global area of the image presentation. The global representations [14, 15] for an image are done based on the comparison of the entire images or entire image windows. Such approaches are well-suited to construct features based on the surrounding factors of the image that include that the entire region for the image is well covered. Besides, another feature type is the local representation [16] for an image. This local representation for an image is done based on the comparison of certain local structures that can match efficiently between images. These sets of local measurements are extracted based on the set of key points of images that describe the essence of the underlying structure of an image.

Writer identification (WI) is an active on-going research area [10, 11, 12]. The WI main problems nowadays are to extract good features [13] that can represent the writer's characteristics or individualistic by providing good input features to the classification model. The problem mainly focuses on the feature extraction techniques that can be constructed either by local [14, 15] or global features extraction [16]. Local features often contain more information than global features but their constructions result in too many features leading to problems with high dimensionality data that can contribute towards poor performance accuracy rate.

Currently, [17] has proposed two methods for writer identification. These methods are meant for text independent type of features for the Delta encoding, which takes into account the meaning of the words while still maintaining text-independent. The Local Contour Distribution Features (LCDF) is proposed by [18]. This method involves the contour detection technique using an improved Bernsen algorithm in pre-processing. Then, the local fragment feature extraction is done repetitively to find all the edges points and normalize the distributions into LCDF. Several studies analyse the feature vectors transformation by representing them in general feature vector to propose the uniqueness of each feature that is able to achieve high classification accuracy [10, 11, 12, 19, 20, 21].

The motivation of this study has come from the area of handwriting analysis and the problem of high dimensionality data to find the most suitable feature extraction method to represent the author's identification feature. This paper is arranged as follows: Section 2 discusses the proposed methods for Global and Local feature extraction methods. The results analysis and discussion are presented in Section 3. Lastly, the conclusion for this paper is presented in Section 4.

## 2. Global and local feature extraction methods

This study proposed the implementation of the feature extraction procedure using two (2) different feature extraction techniques to extract local and global features. The global feature extraction procedure was conducted using Higher Order United Moment Invariant (HUMI). HUMI is a geometrical feature extraction technique from a fusion formulation of United Moment Invariant (UMI) and Higher Order Scaled-Invariants proposed by [21, 22]. Besides, the local feature extraction procedure was conducted using the Edge-based Directional technique proposed by [20]. This study was conducted to determine the benefit of implementing different feature extraction techniques towards the handwriting images.

### *2.1. Higher-Order United Moment Invariant*

Moment Invariants [23] refers to a recognition of geometrical patterns and alphabetical characters that are independent of position, size and orientation. This method is widely being implemented in image analysis and pattern recognition. A new method of United Moment Invariants (UMI) based on some of Hu's moments has been redefined by [24] for the purpose of shape discrimination. These discriminant shape features are also able to be invariant to scaling, translation and rotation towards region, closed and unclosed boundary.

The HUMI algorithm with improved scale-invariant is shown as the following procedures:

1. Generate geometric moments for

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p \; y^q \; f(x,y)\, dxdy \tag{1}$$

The $f(x, y)$ represents the grey level for the image, that is set to 0 for the background and 1 for the object. The representation of discrete form is

$$m_{pq} = \sum_{i=1}^{L} x_i^p \; y_i^q \tag{2}$$

where L represents the number of pixels belongs to the object.

2. The image is centred at $(x_0, y_0)$ where

$$x_0 = \frac{m_{10}}{m_{00}} \qquad \text{and} \qquad y_0 = \frac{m_{01}}{m_{00}} \tag{3}$$

3. Calculate a set of central moments which are invariant to translation

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(x - x_0\right)^p \left(y - y_0\right)^q f(x,y)\, dxdy \tag{4}$$

The representation of discrete form is

$$\mu_{pq} = \sum_{x} \sum_{y} \left(x - \bar{x}\right)^p \left(y - \bar{y}\right)^q f(x,y) \tag{5}$$

4. The normalization is done to generate normalized central moments that are invariant to scaling by integrating the scale factor of higher-order centralized invariants [22] into united moment invariant [24].

$$\eta_{pq} = \left( \frac{\mu_{20}^{p+1/2}\mu_{02}^{q+1/2}}{\mu_{40}^{p+1/2}\mu_{04}^{q+1/2}} \right) \mu_{pq} \tag{6}$$

5. The generation of the 7-tuple features of [23] is done by using the above $\eta_{pq}$ which are invariant to translation, rotation and scaling.

$$\phi_1 = \eta_{20} + \eta_{02} \tag{7}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{8}$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (\eta_{03} - 3\eta_{21})^2 \tag{9}$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{10}$$

$$\begin{aligned}\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\,[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\,[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{11}$$

$$\begin{aligned}\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ &+ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\end{aligned} \tag{12}$$

$$\begin{aligned}\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\,[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &+ (3\eta_{12} - \eta_{03})(\eta_{21} + \eta_{03})\,[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{13}$$

6. Finally, the eight (8) HUMI features are presented by the formulation below:

$$\begin{aligned}&\theta_1 = \frac{\sqrt{\phi_2}}{\phi_1} \quad \theta_2 = \frac{\phi_6}{\phi_1\phi_4} \quad \theta_3 = \frac{\sqrt{\phi_5}}{\phi_4} \quad \theta_4 = \frac{\phi_5}{\phi_3\phi_4} \\ &\theta_5 = \frac{\phi_1\phi_6}{\phi_2\phi_3} \quad \theta_6 = \frac{\left(\phi_1 + \sqrt{\phi_2}\right)\phi_3}{\phi_6} \quad \theta_7 = \frac{\phi_1\phi_5}{\phi_3\phi_6} \quad \theta_8 = \frac{\left(\phi_3 + \phi_4\right)}{\sqrt{\phi_5}}\end{aligned} \tag{14}$$

Due to the large values of $\phi_i$ , the natural logarithm is applied as follows:

$$for\, i = 1 \;\; to \;\; 7; \qquad \theta_1 \leftarrow \log_{10} \phi_i \tag{15}$$

The implementation of the HUMI feature extraction method to acquire the Global Features for this study is aimed to generate better invariants. The example of the image acquisition and representation based on the moment invariants is shown in Figure 1 that shows the word image MOVE for an example Writer Class 1. The original image has the dimension of 290 x 393 pixels that is transformed into the formulation of $\eta_{pq}$ where x takes the value of 290 while y with 393. The image is constructed into eight (8) feature vectors of $\theta_i$ that represents the image.

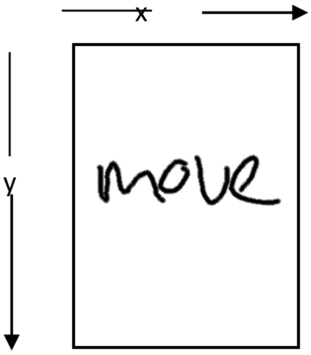


**Figure 1.** Example of Word Image MOVE for Writer Class 1.

**Table 1.** The HUMI Global Features of Word Image MOVE for Writer Class 1.

| $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ | $\theta_6$ | $\theta_7$ | $\theta_8$ |
|---|---|---|---|---|---|---|---|
| 4.6112 | 4.8562 | 0.2565 | 0.5333 | 3.3198 | 5.9026 | 4.3229 | 0.8272 |

The representation and generation of eight (8) HUMI features based on the formulation of $\theta_i$ have given the global feature vectors as shown by Table 1 for the example of word image MOVE for Writer Class 1.

### *2.2. Edge-based directional*

Edge-based Directional feature extraction method [20] was implemented in this study to construct Local Features for writer identification. The extraction of Local Features for this study is aimed to acquire better informative invariant features to determine the writer's identity.

The procedures for Edge-based Directional are as the following:

1. The division of the image into 3 x 3 equal size windows that creates nine (9) individual main zones for the word image.
   a. Each zone is also divided into nine (9) zones for the extraction of feature vectors in each main zones
   b. An extraction of each child zones is based on three (3) types of pixels including the starter points, intersection and minor starter
   c. The traversal of each pixel is based on the neighbouring pixels

2. Determine the line segments for the image.

3. Extraction of features image is done by traversing the nine (9) zones individually.

Finally, each main individual zone constructs the feature vectors.

The same example of word image MOVE given in Figure 2 is used to derive the generation of nine (9) Local Features by Edge-based Directional method. Figure 3 illustrates the 3 x 3 windows that divided the example of the word image MOVE for Writer Class 1 into nine (9) individual zones.

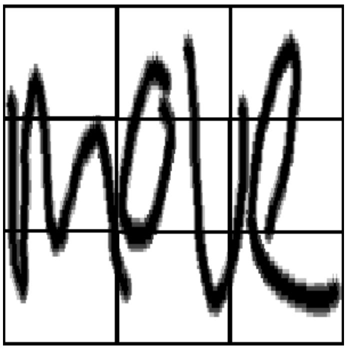


**Figure 2.** Example of Word Image MOVE for Division of Nine (9) Zones

Figure 3 shows the individual nine (9) zones, divided from left window to the right by numbering with Zone 1 until Zone 9 to represent the Local Features extracted by Edge-based Directional method. For a partial image of Zone 4, the feature vectors as shown by Figure 4 that constructed the child zones are produced to determine the individual value of feature vector belonging to Zone 4.

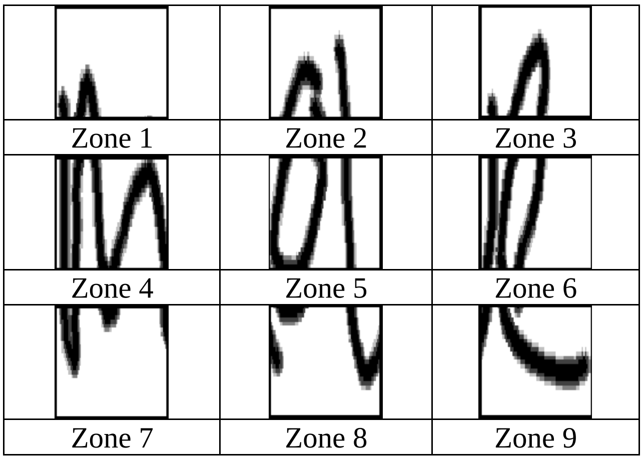


**Figure 3.** Example of Word Image MOVE for Nine (9) Zones of Edge-based Directional for Writer Class 1

| | | | |
|---|---|---|---|
| | 2.8000 | 4.6000 | 1.4000 |
| | 0.0000 | 0.0787 | 0.2677 |
| | 0.0512 | 0.0000 | 0.8789 |
| Zone 4 | Feature Vector: 10.0765 | | |

**Figure 4.** Example of Feature Vectors for Word Image of Zone 4

**Table 2.** The ED Local Features of Word Image MOVE for Writer Class 1.

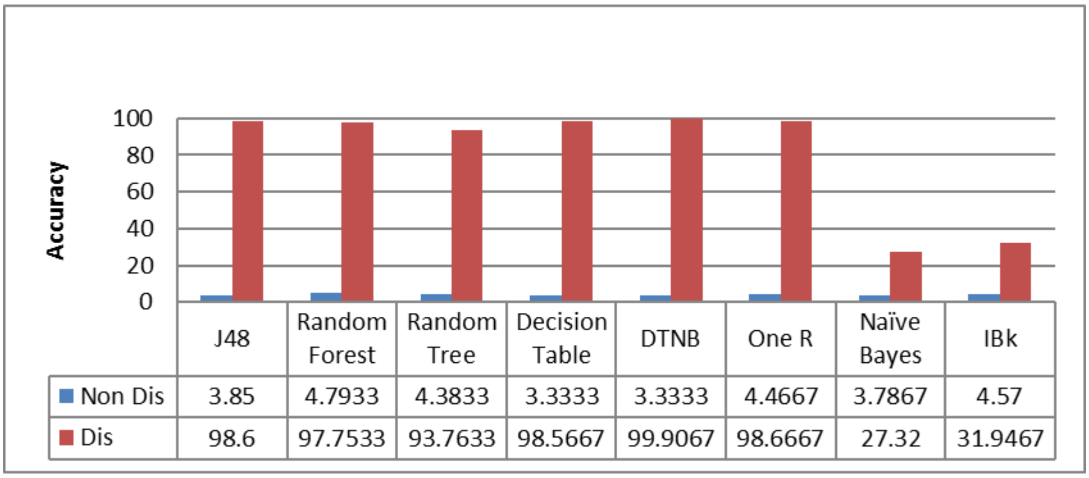
| $z_1$ | $z_2$ | $z_3$ | $z_4$ | $z_5$ | $z_6$ | $z_7$ | $z_8$ | $z_9$ |
|---|---|---|---|---|---|---|---|---|
| 1.0000 | 1.0000 | 0.9412 | 10.0765 | 1.0285 | 0.7958 | 0.9412 | 0.9412 | 0.8858 |

Table 2 shows the example production of nine (9) local feature vectors from the extractions of Edge-based Directional for the same word image MOVE for Writer Class 1.

## 3. Result analysis and discussion

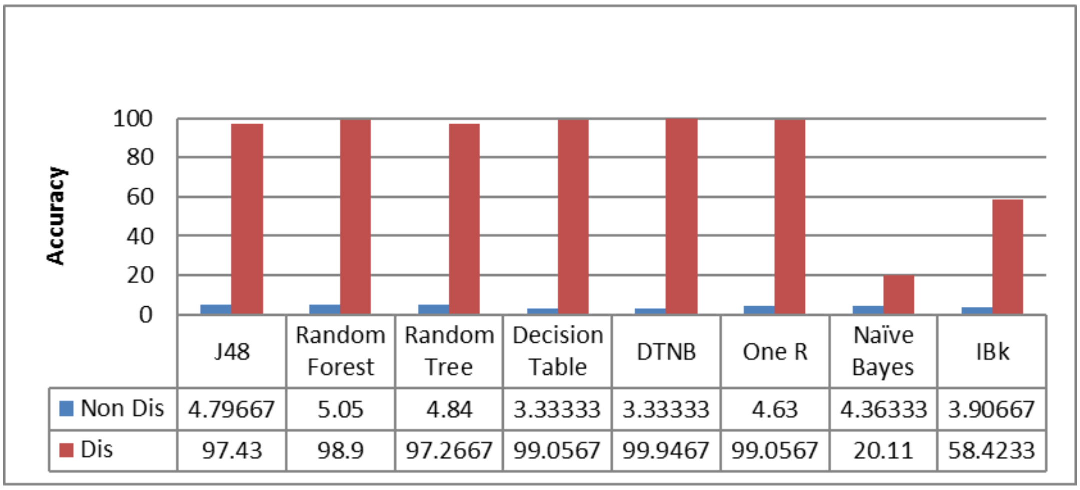
This study analysed the datasets of twenty (20) authors with thirty (30) handwriting word images for each author that comprises of 600 words images altogether. HUMI and ED are used to extract the global and local features, respectively for all the 600 images. The global and local features constructed an eight (8) and nine (9) features respectively for each word image. These features are also being discretized using the supervised discretization method of Equal Width Binning (EWB) [25] to produce better feature representation for the classifier models. The representation of all features is then trained and tested on several classifier models to find their performance.

Figure 5 shows the performance of all discretize and all non-discretized for global features of ten (10) folds cross-validation with fifty (50) runs while Figure 6 shows the performance result for local features with the same settings of the experimental setup.



| | J48 | Random Forest | Random Tree | Decision Table | DTNB | One R | Naïve Bayes | IBk |
|---|---|---|---|---|---|---|---|---|
| 🟦 Non Dis | 3.85 | 4.7933 | 4.3833 | 3.3333 | 3.3333 | 4.4667 | 3.7867 | 4.57 |
| 🟥 Dis | 98.6 | 97.7533 | 93.7633 | 98.5667 | 99.9067 | 98.6667 | 27.32 | 31.9467 |

**Figure 5.** Performance of All Discretize and All Non-Discretized Global Features for ten (10) folds Cross-Validation with fifty (50) runs



| | J48 | Random Forest | Random Tree | Decision Table | DTNB | One R | Naïve Bayes | IBk |
|---|---|---|---|---|---|---|---|---|
| 🟦 Non Dis | 4.79667 | 5.05 | 4.84 | 3.33333 | 3.33333 | 4.63 | 4.36333 | 3.90667 |
| 🟥 Dis | 97.43 | 98.9 | 97.2667 | 99.0567 | 99.9467 | 99.0567 | 20.11 | 58.4233 |

**Figure 6.** Performance of All Discretize and Non-Discretized Local Features for ten (10) folds Cross-Validation with fifty (50) runs

The performance of all discretized local features for DTNB classifier model succeeded with the highest performance of 99.95% accuracy rate, whereas the global features performed slightly lower with 99.91%. This is followed by the classifier model of Decision Table and One R with the performance also above 99% for local features whereas 98% for global features. Besides, Random Forest achieved above 98% with Random Tree, and J48 achieved more than 97% accuracy rate. In comparison with the global features, J48 managed to achieve slightly higher while Random Forest performed at 97% and Random Tree at 93%. Although, two other classifiers achieved quite a lower average performance of 58.42% for IBk and 20.11% for Naïve Bayes for local features, both presented higher and better performance values than all the classifier for all non-discretized features which only able to perform up until 4.84% and almost the same with the global features that achieved 4.79% for all non-discretized features.

This achievement, hence, reflects that discretization procedure for global and local features has contributed to the higher performance rate by representing the features for the classifier models. The results show that the performance accuracy rate for local features is slightly better than the global features. This was due to more discretized local features representing significant and relevant information to identify the authors.

## 4. Conclusion

In conclusion, the comparison performance for both global and local features has been done using discretized and non-discretized features representation. The performance accuracy rate shows that the discretized local features present the highest result of 99.95% that are trained and tested by DTNB classifier. For all other classifiers also, the discretized local features manage to outperform discretized global features except for J48 classifiers when discretized global features achieved slightly higher that is 98% better than discretized local features that perform at 97.43% performance. This shows that the local features extracted by Edge-based Directional (ED) feature extraction method manage to identify the author's handwriting image with slightly better performance than the global features extracted using the Higher-Order United Moment Invariant (HUMI) method with the additional process of discretization to represent generalized features for the classifier models. Although ED presents a slightly better performance for the author's identification, HUMI is also able to give a good performance for most of the classifiers with discretized global features.

**Acknowledgments**
The research is supported by Universiti Teknikal Malaysia Melaka (UTeM) and sponsored by Ministry of Education Malaysia under the Fundamental Research Grant Scheme (FRGS) number FRGS/2018/FTMK-CACT/F00390. 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 & Komunikasi (FTMK), UTeM.

## References

[1] Chandrashekar G and Sahin F 2014 A survey on feature selection methods *Computers and Electrical Engineering* ***40*****(1)** pp 16–28

[2] Test E, Kecman V, Strack R, Li Q and Salman R 2012 Feature ranking for pattern recognition: A comparison of filter methods *Conference Proceedings - IEEE SOUTHEASTCON*

[3] Dittman D, Khoshgoftaar T, Wald R and Napolitano A 2012 Similarity analysis of feature ranking techniques on imbalanced DNA microarray datasets *Proceedings - 2012 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2012* pp 398–402

[4] Inbarani H H, Azar A T and Jothi G 2014 Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis *Computer Methods and Programs in Biomedicine* ***113*****(1)** pp 175–185

[5] Shilaskar S and Ghatol A 2013 Feature selection for medical diagnosis: Evaluation for cardiovascular diseases *Expert Systems with Applications* ***40*****(10)** pp 4146–53

[6] Jiang H and He W 2012 Expert systems with applications grey relational grade in local support vector regression for financial time series prediction *Expert Systems With Applications* ***39*****(3)** pp 2256–62

[7] Song Q and Shepperd M 2011 Expert systems with applications predicting software project effort: A grey relational analysis based method *Expert Systems With Applications* ***38*****(6)** pp 7302–16

[8] Wang H, Khoshgoftaar T M and Gao K. 2010 A comparative study of filter-based feature ranking techniques *2010 IEEE International Conference on Information Reuse & Integration* pp 43-48

[9] Yang S 2013 On feature selection for traffic congestion prediction *Transportation Research Part C: Emerging Technologies* ***26*** pp 160–169

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

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

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

[13] Yang W, Jin L and Liu M 2016 DeepWriterID: An end-to-end online text-independent writer identification system *IEEE Intelligent Systems* ***31*****(2)** pp 45–53

[14] Du L, You X, Xu H, Gao Z and Tang Y 2010 Wavelet domain local binary pattern features for writer identification pp 3695–98

[15] Kumar R, Chanda B and Sharma J D 2014 A novel sparse model based forensic writer identification *Pattern Recognition Letters* ***35*****(1)** pp 105–112

[16] Chaabouni A and Boubaker H 2010 Fractal and multi-fractal for arabic offline writer identification

[17] Newell A J and Griffin L D 2014 Writer identification using oriented basic image features and the delta encoding *Pattern Recognition* ***47*****(6)** pp 2255–65

[18] Ding H, Wu H, Zhang X and Chen J 2014 Writer identification based on local contour distribution feature *International Journal of Signal Processing, Image Processing & Pattern Recognition* ***7*****(1)** pp 169–180

[19] Ali A and Omer B 2016 Invarianceness for character recognition using geo-discretization features *Computer and Information Science* ***9*****(2)** p 1

[20] Leng W Y and Shamsuddin S M 2010 Writer identification for chinese handwriting *International Journal of Advances in Soft Computing & Its Applications* **2(2)** p 142-173

[21] Muda A K, Shamsuddin S M and Abraham A 2009 Authorship invarianceness for writer identification *In 2009 International Conference on Biometrics and Kansei Engineering* pp 34-39 IEEE

[22] Shamsuddin S M, Sulaiman M. N and Darus M 2002 Invarianceness of higher order centralised scaled-invariants undergo basic transformations *International Journal of Computer Mathematics* ***79*****(1)** p 39–48

[23] Hu M K 1962 Visual pattern recognition by moment invariants *IRE transactions on information theory* **8(2)** pp 179-187

[24] Yinan S, Weijun L and Yuechao W 2003 United moment invariants for shape discrimination *IEEE International Conference on Robotics, Intelligent Systems and Signal Processing* **1** pp. 88-93

[25] Muda A K, Shamsuddin S M and Darus M 2008 Invariants discretization for individuality representation in handwritten authorship *International Workshop on Computational Forensics* pp 218-228

[26] Radzid A R, Azmi M S, Jalil I E A, Arbain N A, Draman A K and Tahir A 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)** pp 53-57

[27] Radzid A R, Azmi M S, Jalil I E A, Muda A K, Melhem L B and Arbain N A 2018 Framework of Page Segmentation for Mushaf Al-Quran Based on Multiphase Level Segmentation *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.* **10** pp 028-037