Computer and Information Science; Vol. 9, No. 2; 2016
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Invarianceness for Character Recognition Using Geo-Discretization Features

Aree Ali[1] & Bayan Omer[2]

[1] School of Science, University of Sulaimani, Sulaimani, KRG, Iraq

[2] College of Science and Technology, University of Human Development, Sulaimani, KRG, Iraq

Correspondence: Aree Ali, School of Science, University of Sulaimani, Sulaimani, KRG, Iraq. Tel: 964-751-502-2030. E-mail: aree.ali@univsul.edu.iq

Received: January 28, 2016 Accepted: February 14, 2016 Online Published: March 17, 2016

doi:10.5539/cis.v9n2p1 URL: http://dx.doi.org/10.5539/cis.v9n2p1

Abstract

Recognition rate of characters in the handwritten is still a big challenge for the research because of a shape variation, scale and format in a given handwritten character. A more complicated handwritten character recognition system needs a better feature extraction technique that deal with such variation of hand writing. In other hand, to obtain efficient and accurate recognition rely on off-line English handwriting character, the similarity in the character traits is an important issue to be differentiated in an off-line English handwriting to. In recognizing a character, character handwriting format could be implicitly analyzed to make the representation of the unique hidden features of the individual's character is allowable. Unique features can be used in recognizing characters which can be considerable when the similarity between two characters is high. However, the problem of the similarity in off-line English character handwritten was not taken into account thus, leaving a high possibility of degrading the similarity error for intra-class [same character] with the decrease of the similarity error for inter-class [different character]. Therefore, in order to achieve better performance, this paper proposes a discretization feature algorithm to reduce the similarity error for intra-class [same character]. The mean absolute error is used as a parameter to calculate the similarity between inter and/or intra class characters. Test results show that the identification rate give a better result with the proposed hybrid Geo-Discretization method.

Keywords: document analysis, unique representation, off-line English handwriting, recognition, discretization

1. Introduction

Pattern recognition provides services for various engineering and scientific fields such as computer vision, biology and artificial intelligence. Pattern recognition in handwriting considered as a wide-ranging term which covers all kinds of application field together with identification based on handwriting (Guo, Christian, & Alex, 2010), verification based on handwriting (Srihari, & Ball, 2009), authentication (Muzaffar, & Jurgen, 2009; Behzad, & Mohsen, 2010) and character recognition (Tonghua, Zhang, Guan, & Huang, 2009; Bayan, 2013).

Recently, the field of pattern recognition is considerably improved and revealed due to the emerging applications which are not only challenging but also attracted many researchers' attention. New applications include (data mining, web searching, retrieval of multimedia data, face recognition, handwritten recognition).These techniques require robust and intelligent pattern recognition techniques. Pattern recognition described by (Anil, Robert, & Jianchang, 2000) as a most critical role in human decision making task, even though we as a human can easily refuse to understand how actually human could recognize patterns.

The character recognition based off-line English handwriting is an open research area in pattern recognition and computer vision fields (Bayan, 2013; Binod, & Goutam, 2012). The shape or style in off-line English character is complex and has similarity among some characters (Binod, & Goutam, 2012; Nisha, Hem, & Singh, 2012). However, there are still unique features for each character. These unique features can be generalized as the individual's character handwriting even though there can be complex and high similarity in off-line English language characters. Figure 1 shows an example of off-line English characters and the similarity among them. An improvement step is added to provide a better representation for the input samples from the same or different characters. Extracted features in the feature extraction process show that the character in an off-line English language has similar style or format which affects the accuracy of the performance.

2. Off-Line English Character Individuality

Off-line English Handwriting character has long been considered individualistic and character individuality rests on the hypothesis that each individual character has consistent handwriting (Binod, & Goutam, 2012; Azmi, Kabir, & Badi, 2003; Bayan, 2012; Nisha, Hem, & Singh, 2012). Figure 1 shows the handwriting of the same character and Figure 2 of different character by four writers. Characters are shown as taking a specific texture (Binod, & Goutam, 2012) and can be seen in below figures. The character structure is faintly different for the identical character and completely different for non-identical character, this is known as individuality of English character. Intra-class measurement is showed for features of the same character, and inter-class for different character. Well-being single features must acquire the minimum error of similarity for intra-class and the maximum similarity error for inter-class.

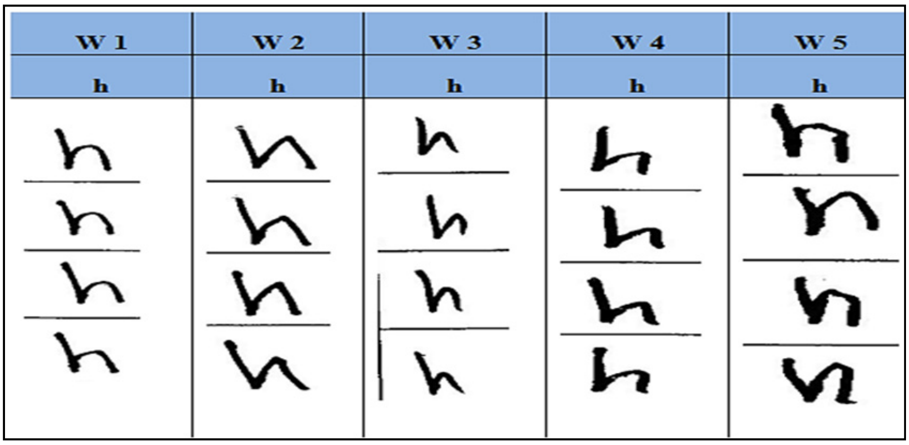

W 1
W 2
W 3
W 4
W 5
h
h
h
h
h

Figure 1. Same Character by Different Writers

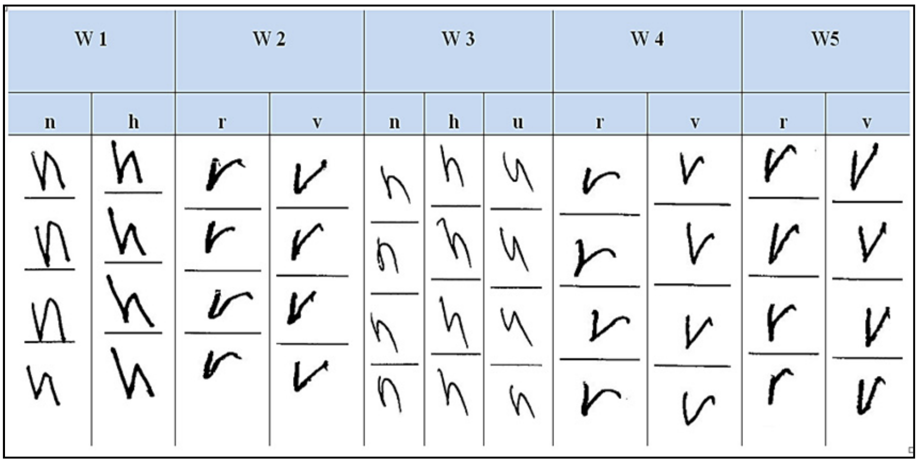

W 1
W 2
W 3
W 4
W5
n
h
r
v
n
h
u
r
v
r
v

Figure 2. Different Character by Different Writers

3. Uniqueness in Off-line English Character Representation

Selecting most predominant features acting as an input to a classifier are very interesting to get better performance in the process of recognition. These kinds of feature do not represent individual features of the character because of representing the character by different features. The proposed method is based on an invariant discretization algorithm which is studying by (Muda, Shamsuddin, & Ajith, 2010; Azmi, Kabir, & Badi, 2003; Bayan, 2012;]. It acts by reducing the dissimilarity between features for intra-class and increasing the dissimilarity between features for inter-class. The traditional and the proposed framework are shown in Figure 3 and 4 respectively.

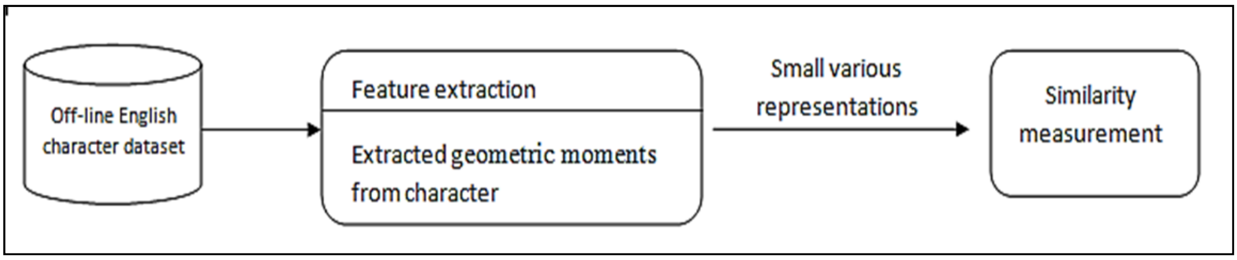

Off-line English character dataset
Feature extraction
Extracted geometric moments from character
Small various representations
Similarity measurement

Figure 3. Traditional Framework

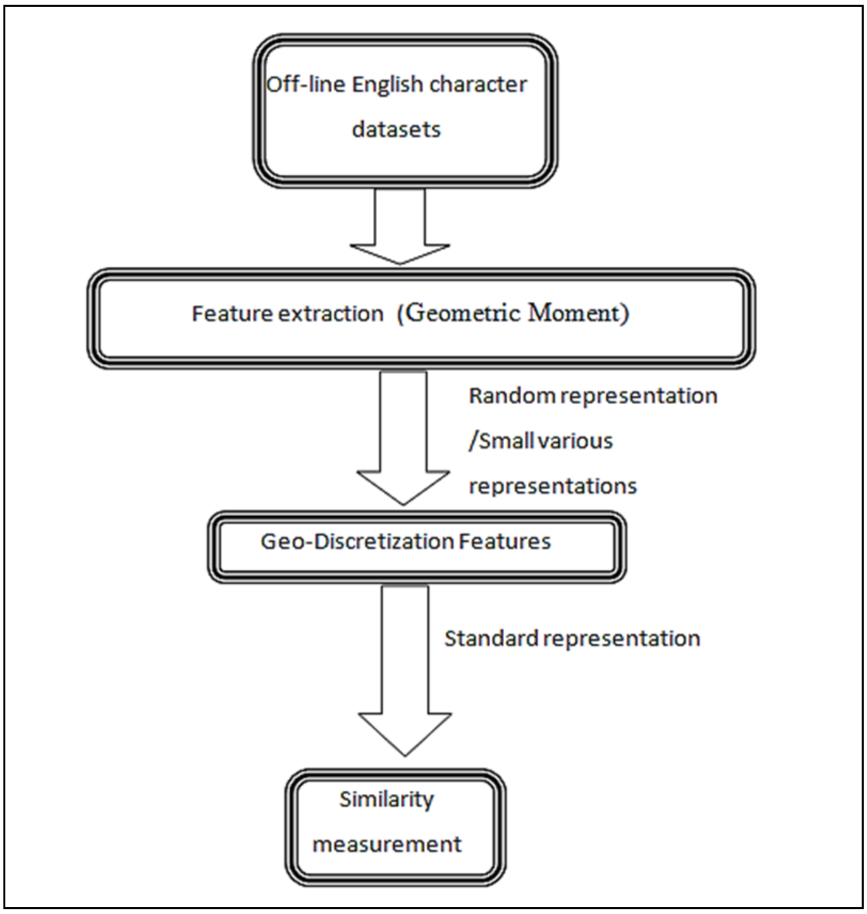

Off-line English character datasets
Feature extraction (Geometric Moment)
Random representation /Small various representations
Geo-Discretization Features
Standard representation
Similarity measurement

Figure 4. Proposed Frame Work Framework

3.1 Discretization Process

Discretization is considered as a divider that performs two essential operations the first task is to convert the value of the continuous characteristics into discrete. The second one is to divide the value and categorized them into appropriate intervals. The main objective of the discretization of the continuous characteristics is to represent the min a better way (Fabrice, & Ricco, 2005). There are some well-known techniques for discretization including Equal Information Gain, Maximum Entropy, and Equal Interval Width. Another method proposed in (Muda, Shamsuddin, & Ajith, Fabrice, & Ricco, 2005)), the Invariants Discretization method, is proved to be better in efficiency by having higher accuracy and better rates of identification. The method is supervised type and starts by choosing the suitable intervals to represent the writer's information (Muda, Shamsuddin, & Ajith, 2010; Fabrice, & Ricco, 2005; Bayan, & Shamsuddin, 2012; Bayan, & Siti, 2011). The upper and lower boundaries are then set for each interval. The number of intervals for an image must be the same as the number of the feature vectors.

3.2 Feature Extraction Phase

Techniques that transform the input sample data into the set of features are called feature extraction method. The characteristic of feature extraction is to reduce the dimension of the given data. Selection of the feature extraction method types is crucial and affects the performance evaluation of any pattern recognition system (Bayan, 2013; Trier, & Jain, 1996). Different extractors are proposed to recognize handwritten digits and characters such as (FT, IM, GM and Characteristic Loci) (Takahashi, 1991; Azmi, Kabir, & Badi, 2003). In this paper, geometric moment method is used to recognize handwritten off-line English characters. Geometric Moment is used in object recognition and pattern recognition applications. A set of distinctive features computed for an object must be capable of identifying the same object with another possible different size and orientation (Muralidharan, & Chandrasekar, 2011; Bayan, 2015; Bayan, 2012).

The computation steps of geometric moments are described as below:

1) Read an input image data from left to right and from top to bottom.

2) Threshold the image data to extract the target process area.

3) Compute the image moment value, m_{pq} until third order with formula:

$$m'_{pq} = \int \int_{\delta} (x')^{p}(y')^{q} f'(x', y') dx'dy' \quad ; \quad p, q = 0,1,2, \ldots . \tag{1}$$

4) Compute the intensity moment, $(x_0, y_0\ )$ of image with formula:

$x_0 = m_{10}/m_{00}\,; \quad y_0 = m_{01}/m_{00}$ (2)

5) Compute the central moments, μ_{pq} with formula:

$$\mu_{pq} = \iint_{\delta} (x - x_0)^{p}(y - y_0)^{q} f(x, y) dxdy \ ; \quad p, q = 0,1,2 \ldots \tag{3}$$

6) Compute normalized central moment, $\eta_{pq}$to be used in image scaling until third order with formula:

$$\gamma = \frac{(p + q + 2)}{2}, \qquad \eta_{pq} = \frac{\mu_{pq}}{\frac{\gamma}{\mu_{00}{}^{2}}} \quad , \qquad p + q \leq 3 \tag{4}$$

7) Compute geometric moments, ϕ_1 to$\phi_4$with respect to translation, scale and rotation (geometric moment invariants) invariants with formula below:

$$\phi_1 = \eta_{20} + \eta_{02} \tag{5}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + \ 4\,\eta_{11}^2 \tag{6}$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{7}$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^{\,2} \tag{8}$$

3.3 Discretization Phase

The process of discretization determines a set of interval that shows the representation of features to be extracted. To obtain an interval, the lowest and highest data range of every writer is distributed along number of intervals (cuts) with equally size. Interval numbers are described according to the number of feature vector in the feature extraction process. An interval value representation is estimated based on the character class. If two characters have an identical invariant value, they take identical interval for these two classes. The Discretization method does not affect or change the properties of character; it is only representing the basic feature vector which is extracted invariantly in a standard representation with global features. Figure 5 depicts the discretization process method.

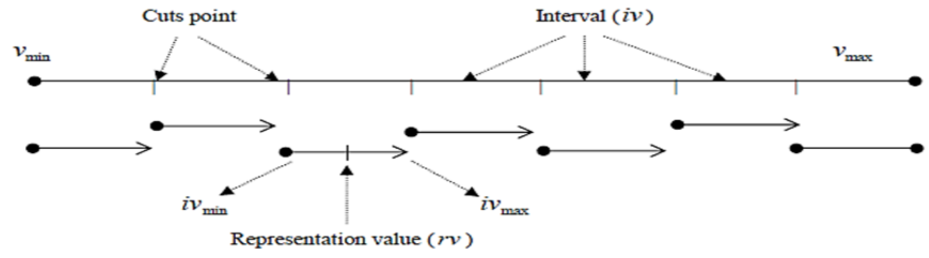


Figure 5. Invariant Discretization Line (Muda, Shamsuddin, & Ajith, 2010)

Invariant discretization line uses minimum ($v_{\min}$) and the maximum ($v_{\max}$) feature vectors to determine the invariant intervals range. The width of an interval can be found as:

$$\text{Width}= (V_{\max}\text{-}\ V_{\min})/f \quad , \tag{9}$$

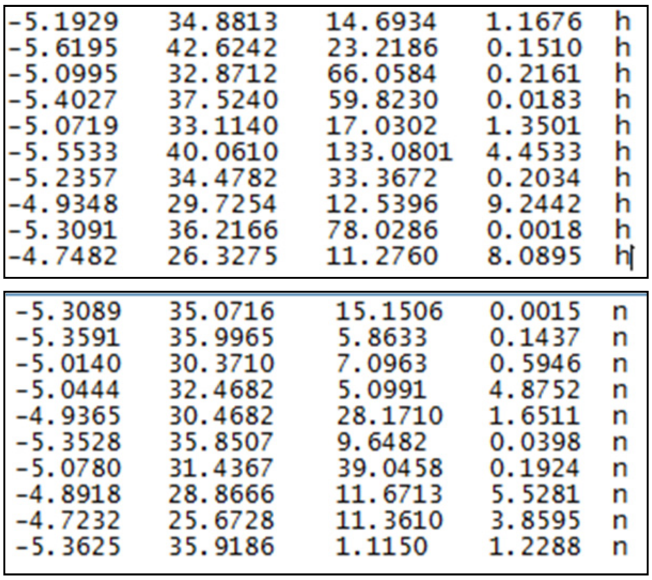
Where:

$v_{\min}$: represent lowest value for a character.

$v_{\max}$: represent highest value for a character.

f: represent invariant feature vector number.

The width in equation (9) is performed to find out the number of cut points of in the discretization line process. Figure 6 and 7 illustrate the process of transformation from invariant feature vector to the discretized feature vector respectively.

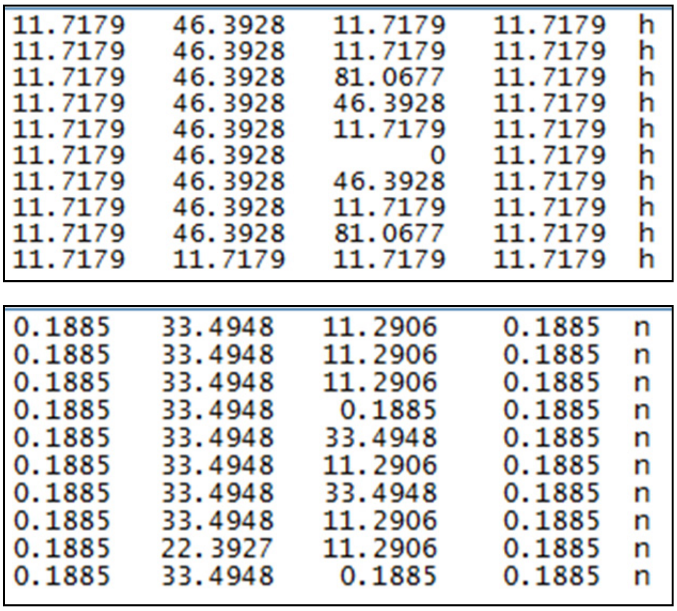


-5.1929	34.8813	14.6934	1.1676	h
-5.6195	42.6242	23.2186	0.1510	h
-5.0995	32.8712	66.0584	0.2161	h
-5.4027	37.5240	59.8230	0.0183	h
-5.0719	33.1140	17.0302	1.3501	h
-5.5533	40.0610	133.0801	4.4533	h
-5.2357	34.4782	33.3672	0.2034	h
-4.9348	29.7254	12.5396	9.2442	h
-5.3091	36.2166	78.0286	0.0018	h
-4.7482	26.3275	11.2760	8.0895	h

-5.3089	35.0716	15.1506	0.0015	n
-5.3591	35.9965	5.8633	0.1437	n
-5.0140	30.3710	7.0963	0.5946	n
-5.0444	32.4682	5.0991	4.8752	n
-4.9365	30.4682	28.1710	1.6511	n
-5.3528	35.8507	9.6482	0.0398	n
-5.0780	31.4367	39.0458	0.1924	n
-4.8918	28.8666	11.6713	5.5281	n
-4.7232	25.6728	11.3610	3.8595	n
-5.3625	35.9186	1.1150	1.2288	n

Figure 6. Invariant Feature Vector Data for Character (h) and (n)

The discretized data yielded from the discretization scheme clearly shows the unique feature of every character in English handwriting.



11.7179	46.3928	11.7179	11.7179	h
11.7179	46.3928	11.7179	11.7179	h
11.7179	46.3928	81.0677	11.7179	h
11.7179	46.3928	46.3928	11.7179	h
11.7179	46.3928	11.7179	11.7179	h
11.7179	46.3928	0	11.7179	h
11.7179	46.3928	46.3928	11.7179	h
11.7179	46.3928	11.7179	11.7179	h
11.7179	46.3928	81.0677	11.7179	h
11.7179	11.7179	11.7179	11.7179	h

0.1885	33.4948	11.2906	0.1885	n
0.1885	33.4948	11.2906	0.1885	n
0.1885	33.4948	11.2906	0.1885	n
0.1885	33.4948	0.1885	0.1885	n
0.1885	33.4948	33.4948	0.1885	n
0.1885	33.4948	11.2906	0.1885	n
0.1885	33.4948	33.4948	0.1885	n
0.1885	33.4948	11.2906	0.1885	n
0.1885	22.3927	11.2906	0.1885	n
0.1885	33.4948	0.1885	0.1885	n

Figure 7. Example of Discretized Feature Data for Character (h)and(n)

4. Uniqueness Test Results

Mean Absolute Error (MAE) function is used to measure the uniqueness of the character. Table. 1 and 2 present the test result values of the MAE when the number of samples is 10 for every character. Feature (1 to 4) is an extracted feature that represents a character. The invarianceness of character and reference image (first image) is given by the MAE value. The small errors mean that the image is close to the reference image. An average of MAE is taken from the value of whole results.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{f}|(x_i - r_i)| \tag{10}$$

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.

Table 1. MAE Results using Geometric Moments

Image	Feature 1	Feature 2	Feature 3	Feature4	MAE
	-5.8208	43.3300	145.7606	10.6865	-
	-5.2686	34.0231	32.6906	0.0387	13.3577
	-4.8517	29.3801	51.8342	4.1312	11.5401
	-5.5301	37.6985	83.2967	6.1862	7.2886
	-5.6034	39.7011	139.7753	10.0286	1.0490
	-5.0146	29.7276	5.0744	0.3166	16.5465
	-5.2040	34.7365	31.9852	0.6482	13.3024
	-5.3170	35.3809	38.1249	0.0696	12.6706
	-5.2492	36.7666	150.5069	7.6241	1.4944
	-5.1119	31.3749	5.8567	0.3519	16.2903
		Average MAE			9.353

Table 2. MAE Results using Geo-Discretization

Image	Feature 1	Feature 2	Feature 3	Feature4	MAE
	13.7202	52.8021	130.9659	13.7202	--
	13.7202	52.8021	13.7202	13.7202	11.7246

	13.7202	13.7202	52.8021	13.7202	11.7246
	13.7202	52.8021	91.8840	13.7202	3.9082
	13.7202	52.8021	130.9659	13.7202	0
	13.7202	13.7202	13.7202	13.7202	15.6328
	13.7202	52.8021	13.7202	13.7202	11.7246
	13.7202	52.8021	52.8021	13.7202	7.8164
	13.7202	52.8021	130.9659	13.7202	0
	13.7202	13.7202	13.7202	13.7202	15.6328
		Average MAE			7.8164

The profession of writing invarianceness for the geometric (moment and Geo-discretized) data value is determined by applying the intra-class and inter-class analysis of MAE value. The test result demonstrates that the dissimilarity between feature for intra-class (identical character) and inter-class (non-identical character) using the Geo-Discretization scheme gives a better result compared to geometric moments data. It has improved the recognition process where the MAE value for intra-class using Geo-discretized data is smaller than geometric moment's data, and MAE value for inter-class using Geo-discretized data is higher than geometric moment's data. The minimum MAE value in intra-class indicates that features are highly identical to each other for the identical character whilst the maximum MAE value for inter-class indicates that they are widely differ to each other for non-idnetical characters. These results have proved the hypothesis that the discretization process can improve the recognition process with a standard representation of individual features for the individuality representation in off-line English handwriting character. Figure 8 and 9 show the MAE results comparison of recognition process for the Geometric feature technique with Geo-discretized data and geometric moment's data.

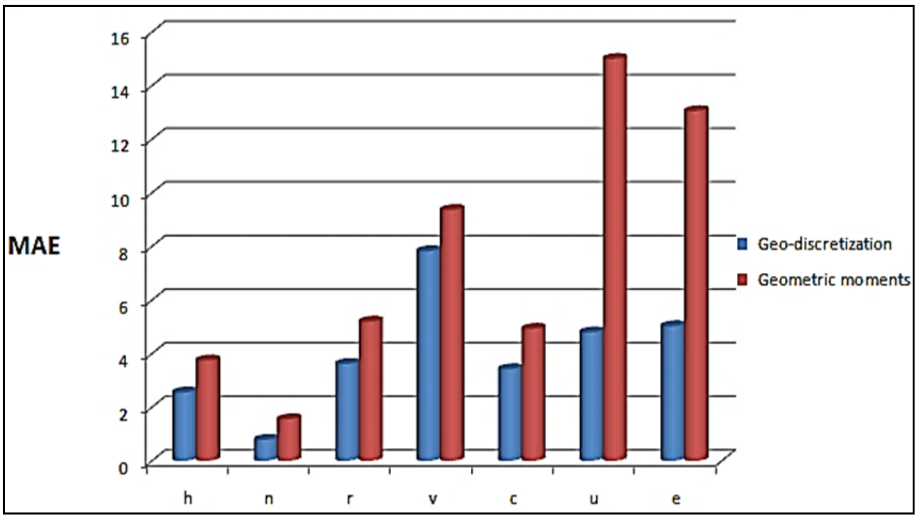


Figure 8. MAE Comparison for Intra-Class

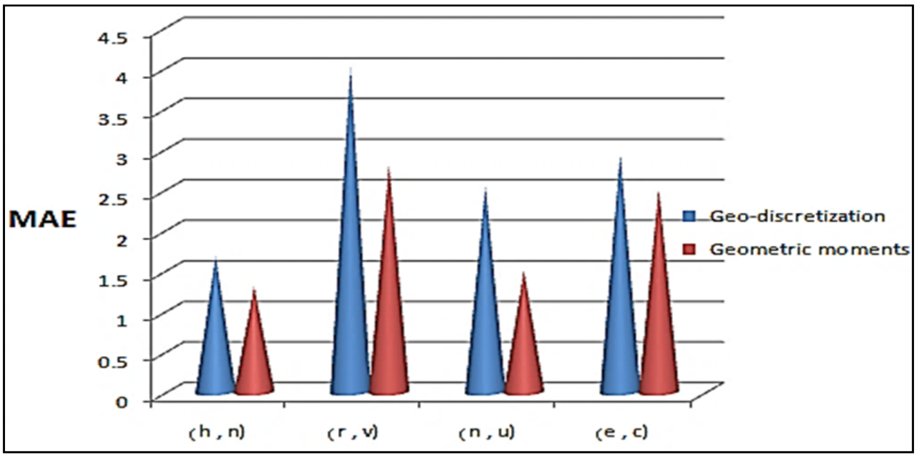


Figure 9. MAE Comparison for Inter-Class

5. Discussion

In this research work, a new framework for off-line English Handwritten Character Recognition is proposed. The effect of discretization process is shown during successful experimental tests. Individual features in the handwritten character can be systematically represented with the use of the invariants discretization algorithm. The results reveal that with the use of the invariant discretization technique, the accuracy of the off-line English handwritten character recognition is significantly improved with the general arrangement to get improved accuracy paralleled to geometric moment's information. For the future work, the similar experiment could be done over some other characters to improve more the accuracy of the proposed method.

References

Anil, K. J., Robert, P. W., & Jianchang, D. M. (2000). *Statistical pattern recognition: A review*. In Proc. 4th IEEE Trans on Pattern analysis and Machine intelligence, 22, 4-37.

Azmi, K., Kabir, R., & Badi, E. (2003). Recognition printed letters witZonong features. *Iran Computer Group,* (1), 29-37.

Bayan, O. M. (2015). Individuality Representation in Character Recognition. *Journal of University of Human Development, 1*(2), 300-305.

Bayan, O. M., & Shamsuddin, S. M. (2012). Improvement in twins handwriting identification with invariants discretization. *EURASIP Journal on Advances in Signal Processing, 48,* 3-12. http://dx.doi.org/10.1186/1687-6180-2012-48

Bayan, O. M., & Siti, M. S. (2011). Feature discretization for individuality representation in twins handwritten identification. *Journal of Computer Science, 7*(7), 1080-1087.

Bayan, O. M. (2012). Uniqueness in Kurdish handwriting. *International Journal of Engineering & Computer Science IJECS-IJENS, 12*(6), 42-50.

Bayan, O. M. (2013). Handwritten Kurdish character recognition using geometric discretization feature. *International Journal of Computer Science, 4*(1), 51-55.

Behzad, H., & Mohsen, M. (2010). A text-independent Persian writer identification based on feature relation graph (FRG). *Pattern Recognition,* (43), 2199–2209. http://dx.doi.org/10.1016/j.patcog.2009.11.026

Binod, K. P., & Goutam, S. (2012). A model approach to off-line English character recognition. *International Journal of Scientific and Research Publications, 2*(6), 1-6.

Fabrice, M., & Ricco, R. (2005). Discretization of continuous attributes. In John Wang (Ed.), *Encyclopedia of Data Warehousing and Mining* (pp. 397-402).

Guo, X. T., Christian, V. G., & Alex, C. K. (2010). Individuality of alphabet knowledge in online writer identification. IJDAR Springer Berlin. *Heidelberg, 13*(2), 145-147. http://dx.doi.org/10.1007/s10032-009-0110-z

Muda, A. K., Shamsuddin, S. M., & Ajith, A. (2010). Improvement of authorship invarianceness for individuality representation in writer identification. *Neural Network World, 3*(10), 371–387.

Muralidharan, R., & Chandrasekar, C. (2011). Object Recognition using SVM-KNN based on geometric moment invariant. *International Journal of Computer Trends and Technology, 1*(3), 215-219.

Muzaffar, B., & Jurgen, K. (2009). *Person authentication with RDTW based on handwritten PIN and signature with a novel biometric smart Pen device*, IEEE Workshop, 63-68.

Nisha V., Hem J. P., & Singh V. (2012). Offline character recognition system using artificial neural network. *International Journal of Machine Learning and Computing, 2*(4), 449-452.

Srihari, N. S., & Ball, R. G. (2009). *Semi-supervised learning for handwriting recognition*. ICDAR, 26-30.

Takahashi H. (1991). *A neural Net OCR using geometrical and zonal pattern features*. In Proc. 1th. Conf. Document Analysis and Recognition, 821-828.

Tonghua, S., Zhang, T. W., Guan, D. J., & Huang, H. J. (2009). Off-line recognition of realistic chinese handwriting using segmentation-free strategy. *Pattern Recognition*, *42*(1), 167-182. http://dx.doi.org/10.1016/j.patcog.2008.05.012

Trier, I. D., & Jain, A. K. (1996). Feature extraction methods for character recognition: A survey. *Pattern Recognition, 29*(4), 641- 662. http://dx.doi.org/10.1016/0031-3203(95)00118-2

Copyrights