Individuality Representation in Character Recognition

Bayan Omar Mohammed
Department of Computer Science,College of Science and Technology ,
University of Human Development,
Kurdistan Region, Iraq
bayancomputer@yahoo.com

Abstract

The task of recognition that is based on handwriting characters in the Kurdish language is an interesting study in the area of computer vision and pattern recognition. In the past couple of years, numerous state-of-the-art techniques and methods have been created for pattern recognition. On the other hand, Kurdish language handwriting recognition has been seen to be more difficult when compared to other different languages. The similarities in the properties in Kurdish characters is the primary reason of the great resemblance in the features of Kurdish handwriting characters, therefore the requirement for the recognition process is critical. Consequently, to obtain accurate and precise recognition on the basis of the Kurdish handwriting character, it is crucial for the resemblances in the character properties of Kurdish handwriting to be distinguished. To identify a particular character, the style of character handwriting may be evaluated to enable the implied representation of the hidden unique features of the user's character. Unique features may guide in recognizing characters that may be important when recognizing the correct character among similar characters. On the other hand, the problem of the resemblances in the properties of handwriting of Kurdish characters were not taken into account ,consequently leaving a high chance of reducing the similarity error for any intra-class (of the same character),with the reduction of the similarity error for any inter-class (of different characters) as well. In order to obtain higher effectiveness, this study uses discretization features for reducing the similarity error for intra-class (of the same character),with the increase of the similarity error for inter-class (of different characters)in recognition of Kurdish Handwriting characters with MAE.

1. INTRODUCTION

General, pattern recognition is a crucial portion of different scientific and engineering areas such as artificial intelligence, biology and computer vision. Pattern recognition of handwriting is assumed a broad-ranging term that is involved in all forms of application areas along with identification process based on handwriting [1], verification process based on handwriting [3], authentication process [2, 5] and character recognition process [4,6].

Currently, focus on the field of pattern recognition is increasing at a fast rate because of the newer applications that are not only a challenge, but also attract the attention of a numerous amount of researchers. These new applications include web searching, data mining, retrieval of multimedia, handwritten recognition, face recognition, which require intelligent pattern recognition methods. Pattern recognition is explained by [7] as the most important role in decision making tasks by humans, even though humans may simply reject to comprehend how humans may recognize certain patterns.

Kurdish handwriting character recognition is a well known field of research within the computer vision and pattern recognition and fields, because in certain scenarios, it gives the only way of identifying the actual character of a written text out of a set of character [6,9]. The structure or general style of a Kurdish character is not simple, and there is similarity between several of the characters in the Kurdish language [9]. On the other hand, there remain specific and unique features for every character. These specific and unique features may be generalized as a person's character handwriting, even though there may be complicated and great similarity in the characters of the Kurdish language. Figure 2depicts an example of the resemblance between Kurdish characters.

On the other hand, a great deal of academic work did not focus on the extra stage that this study considers. The extra stage seeks to give a more accurate representation for the input that is used in the process of MAE. A more detailed representation of the input may help in a strategy where the MAE process may be complete faster and more effectively for the actual character to be more effectively identified, specifically in the scenario of Kurdish character recognition. The features retrieved in the feature extraction task reveal that Kurdish language characters have many similar representations, and this causes an issue when the input is implemented in the MAE task as resemblances will decrease the effectiveness of the performance process. This study gives a discussion on the extra stage of transformation where the very similar representation of features are changed into more clear and better representations that may represent every character in the Kurdish language.

2. Individuality of Kurdish Handwriting character

Handwriting of Kurdish characters has been assumed individualistic, and the individuality of characters is based on the theory that every single character has handwriting that is consistent [9]. Fig. 1 depicts the handwriting of the characters that are the same and Fig. 2 depicts the handwriting of characters that are different by four different individuals. Every single character is assumed to have a particular texture [9] and maybe witnessed in the two figures. The general shape partially differs for the exact same character and somewhat varies for other characters. This is referred to as the individuality of characters in the Kurdish language. The intra-class estimation and measurement is shown for features of the same exact character, and inter-class for different characters. Suitable unique features have to generate the least similarity error for intra-class and the greatest similarity error for inter-class. As a result, it is important to obtain unique features from a character to overcome this issue for the recognition of handwritten characters in the Kurdish language.

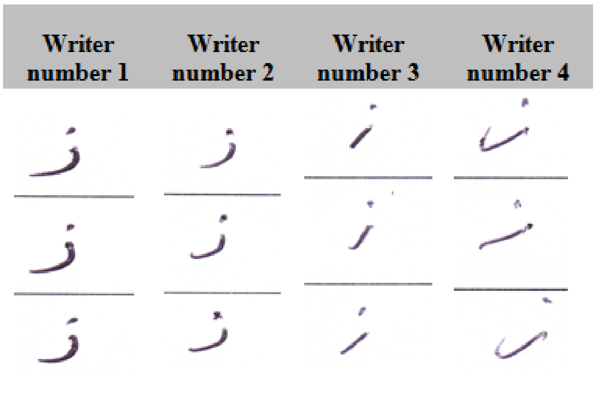

Fig. 1.Same characters written by several individuals

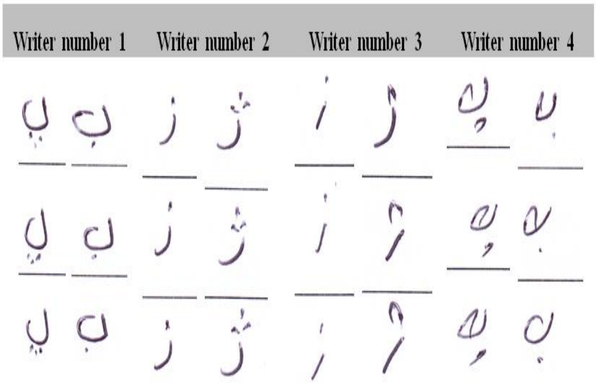

Fig. 2.Different characters written by several individuals

3. Individuality Representation

Suitable features that are input to a classifier are crucial to achieve high effectiveness in the task of recognition. Commonly, retrieved features directly undergo the recognition process to identify a particular character. These features do not represent the unique features of some character among characters of the Kurdish language because the character handwriting in Kurdish possesses similar features, and this leads to a slight different in the handwriting among characters in Kurdish. Commonly, a different task is required to enhance the identification performance. This proposed task is referred to as invariant discretization, studied by [8] and it is aimed to decrease the variance among features for intra-class and increase the variance among features for inter-class. A general outline of a novel framework that is required as an extra process before the measurement of similarity task to enhance the effectiveness of the recognition of Kurdish characters. The conventional framework is depicted in Fig. 3,and the new framework is depicted in Fig. 4.

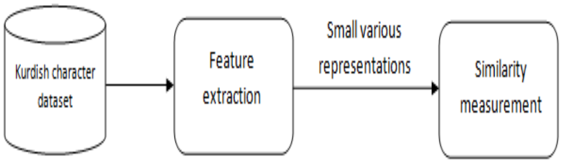

Fig. 3.Traditional framework [12]

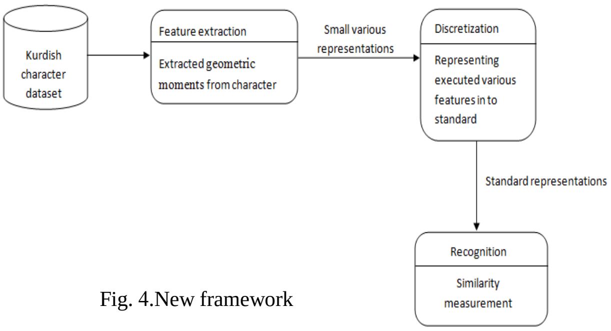
Kurdish character dataset
Feature extraction
Extracted geometric moments from character
Small various representations
Discretization
Representing executed various features in to standard
Standard representations
Recognition
Similarity measurement

Fig. 4.New framework

4. Discretization Process

The process of discretization is acting as a separator that carries out two primary steps. The first step is to transform the overall value of the existing continuous characteristics into discrete. The second step is to separate the overall value and classify them into suitable intervals. The primary reason of the discretization of the continuous characteristics is to generate a more accurate representation of data [10]. Generally, classification greatly depends on the process of discretization. There are a number of common techniques for discretization, such as Equal Information Gain, Equal Interval Width and Maximum Entropy. A different technique is proposed in [8], the Invariants Discretization technique, has been seen to be faster in giving greater accuracy and success rates of identification. The Invariants

Discretization technique is a supervised one. The technique is initialized by seeking the suitable intervals to symbolize information an individual [8,10,11,12]. Both the lower and upper boundaries are then created for every interval. The amount of intervals for a particular image has to be equal to the amount of feature vectors.

5. Adaptation in Kurdish handwriting character recognition

This section explains adaptations of the solution that is propose in showing unique

features and enhancing the difference among features for both intra-class and inter-class in Kurdish character recognition .

5.1 Feature extraction

The transformation of the input data into a group of features is referred to as features extraction. Generally feature extraction is a unique characteristic of dimensionality reduction techniques. Analysis with a great amount of variables needs a great deal of computational power and memory or a general classification algorithm, which over fits the sample used for training and does not perform well with samples that are new. If the data input is very large and cannot be processed, the input data is changed into a decreased representation group of features. Clearly, it is critical to choose a kind of feature extraction technique, because it is a very important factor in the overall performance value of pattern recognition systems [6,13]. The choosing of the kind of feature extraction greatly relies on the application. Various features are aimed to identify written characters and numbers. They are Furies Transform, Characteristic Loci, Invariant Moments and Geometric Moments [14,15]. In this study, the researchers use Geometric Moments identify Kurdish characters that are handwritten. Geometric Moment is implemented in recognition of objects and pattern recognition applications. A group of unique features calculated for a certain object have to be able to identify the same object with a different orientation and size [16].

The steps of computation of geometric moments are explained as the following [6,16]:

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

2) Threshold all the image data to retrieve the target process location.

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

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

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

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

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

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

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

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

7) Compute geometric moments, Φ_1 0 to 0Φ_4 with respect to translation, scale and rotation (geometric moment invariants) invariants with formula below:

$$\Phi_1 = \eta_{20} + \eta_{02} \qquad (5)$$

$$\Phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \qquad (6)$$

$$\Phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \qquad (7)$$

$$\Phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \qquad (8)$$

5.2 Invariants Discretization process

An appropriate group of intervals to symbolize the retrieved features with a certain representation is computed in the process of discretization. This certain representation value is referred to as a discretized feature vector, and the “generalized and unique feature" of unique features is retrieved from the median value of an interval. This generalized feature is implemented to show the unique feature that is not shown in the unique character of handwriting in Kurdish. In order to obtain an interval, the values of maximum and minimum data of every writer is separated into a set of intervals that have the same size. The amount of intervals is given according to the amount of feature vector columns in the retrieved features. In the example, a total of four feature vector columns are retrieved from the geometric feature method. Upper and lower approximation is assigned to the every interval, and every one of the intervals is shown by a single representation value. The invariant feature vector that lies in the same interval also has a similar representation. The representation value for an interval is computed according to the character class (with supervised discretization). If any two characters have an invariant feature vector that is similar, they will have exactly the same or very similar intervals (or cuts) for the two classes. The discretization algorithm keeps the characteristics and properties of characters the same. It represents the main retrieved invariant feature vector in a general representation along with generalized features. Fig. 5 shows the discretization algorithm.

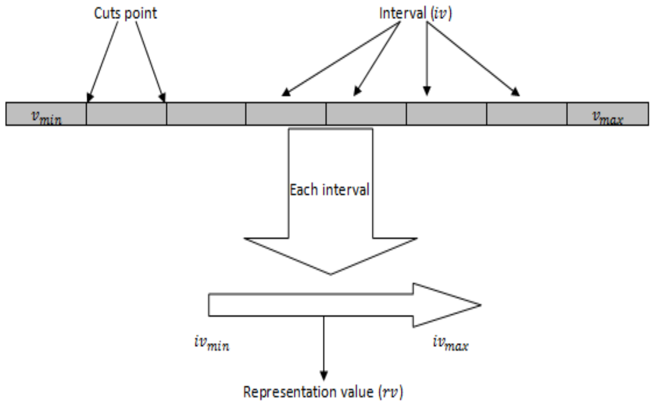

Fig. 5.Invariant Discretization Line [8]

As previously mentioned, the invariant discretization requires the character class information in the process of discretization. The minimum (v_{min}) and the maximum (v_{max}) values of invariant feature vectors (if v) for a certain character are implemented to compute the intervals within the invariant discretization line. The line initiates from the minimum (v_{min}) invariant feature vector and stops with the maximum (v_{max}) invariant feature vector value for a certain character. Generally an interval is the average of the invariant discretization line separated by the total amount of invariant feature vector columns. The width (or wd) of an interval may be computed with:

$$wd = (V_{max} - V_{min})/f \quad (9)$$

where:

v_{min}: The minimum value of invariant feature vector for any character.

v_{max}: The maximum value of invariant feature vector for any character.

f: The total number of invariant feature vector columns. The width is also used to give a definition to cut points of any interval within the invariant discretization line. Every invariant feature vector that lies in the same interval has an identical representation value. The representation value (rv) for every interval is generally the average of interval that is computed with the use of $rv= (iv_{max} - iv_{min})/2$. The general representation value for the intervals of one through four symbolize the invariant feature vector that lies in the range of if $v \geq iv_{min}$ and if $v < iv_{max}$. Also, the invariant feature vector lies within if $v \geq iv_{min}$ and if $v \leq iv_{max}$ is classified as the last interval. This general representation value is referred to as discretized feature vector that symbolizes the unique features of any character. A simple example of the change of invariant feature vector into a discretized feature vector is shown in Fig. 6 and Fig. 7.

-4.598	26.026	33.258	17.90	ز
-4.801	28.135	18.657	7.674	ز
-4.954	29.513	20.737	2.490	ز
-5.075	31.709	25.992	2.527	ز
-5.150	32.742	23.574	2.456	ز
-4.939	29.602	44.419	5.046	ز
-5.084	32.138	10.387	1.657	ز
-4.876	28.325	25.800	4.452	ز

-4.831	28.611	73.380	10.046	ژ
-5.333	35.470	139.255	0.204	ژ
-1.673	14.780	896.269	270.69	ژ
-5.017	30.968	37.253	1.046	ژ
-4.903	29.062	11.225	10.441	ژ
-4.783	26.383	67.272	3.965	ژ
-4.940	29.142	19.680	8.627	ژ
-5.141	32.620	125.106	1.474	ژ

Fig. 6. Invariant Feature Vector Data for Character (ژ) and (ز)

The discretized feature vector retrieved from the discretization algorithm shows the unique features of every character in handwriting in the Kurdish language.

1.6158	28.7347	28.7347	15.1752	ز
1.6158	28.7347	15.1752	1.6158	ز
1.6158	28.7347	15.1752	1.6158	ز
1.6158	28.7347	28.7347	1.6158	ز
1.6158	28.7347	28.7347	1.6158	ز
1.6158	28.7347	42.2942	1.6158	ز
1.6158	28.7347	15.1752	1.6158	ز
1.6158	28.7347	28.7347	1.6158	ز

107.3673	107.3673	107.3673	107.3673	ژ
107.3673	107.3673	107.3673	107.3673	ژ
107.3673	107.3673	107.3673	332.7677	ژ
107.3673	107.3673	107.3673	107.3673	ژ
107.3673	107.3673	107.3673	107.3673	ژ
107.3673	107.3673	107.3673	107.3673	ژ
107.3673	107.3673	107.3673	107.3673	ژ
107.3673	107.3673	107.3673	107.3673	ژ

Fig. 7. Example of Discretized Feature Data for Character (ژ) and) (ز)

5.3 Uniqueness in Kurdish handwriting character

The uniqueness is estimated with the use of the Mean Absolute Error (MAE) function. An example of the MAE computation is shown in Tab. I. There are a total of 15 images for every character. Feature 1 to Feature 4 is a retrieved feature that symbolizes a particular character. The invariance of that character and the reference image (the first image) is supplied by the MAE value[8]. The small errors show that the image is near the corresponding reference image. An average value of MAE is extracted from the value of total results.

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

where,

n is the amount of images.

x_i is the image in focus.

r_i is the location measure or reference image.

f is the a mount of features.

i is the feature column of a certain image.

TABLE I. EXAMPLE OF MAE CALCULATION

Image	Featur-e 1	Featur-e 2	Featur-e 3	Featur-e 4	MAE
	-4.8310	28.6110	73.3800	10.0460	-
	-5.3330	35.4700	139.2550	0.2040	5.538
	-5.1410	32.6200	125.1060	1.4740	4.307
	-4.9730	29.9170	42.2800	1.9190	2.711
	-5.1350	31.7670	62.9830	0.6630	1.549
	-4.6510	26.5560	25.5290	19.3990	3.962
…		…	….	…	…
…		…			
	-4.7830	26.3830	67.2720	3.9650	0.964
Average MAE					7.870

The authorship invariance for a certainpre-discretized feature vector and a certain post-discretized feature vector is computed by conducting both the intra-class and inter-class analysis of the MAE value. The general analysis result depicts that the general variance between the intra-class feature (the same character) and the inter-class feature (different character) using the post-discretized feature vector yields a more appropriate outcome when compared to pre-discretized data. It has enhanced the process of recognition where the MAE value for intra-class implementing post-discretized data is less than pre-discretized data, and the MAE value for inter-class implementing post-discretized data is greater than pre-discretized data. The least MAE value in intra-class shows that the features are very similar to one another for the exact same character while the greatest MAE value for inter-class shows that they are very different from one another for characters that are different. These results have verified the hypothesis that the process of discretization may enhance the overall process of recognition with a standard representation of unique features for the individuality representation in handwriting in the Kurdish language. Fig. 8 and Fig. 9 depict a general comparison of the recognition process for the geometric feature method with both post-discretized data and also pre-discretized data with MAE.

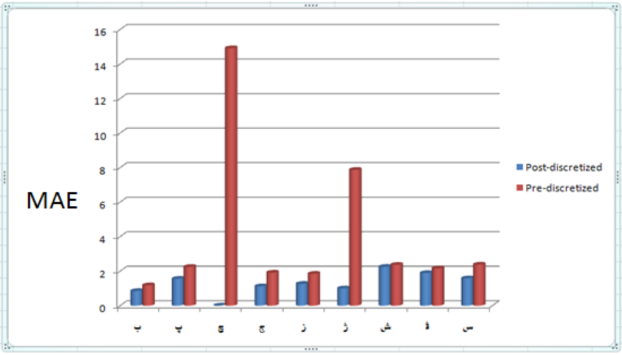

Fig. 8. MAE comparison for intra-class

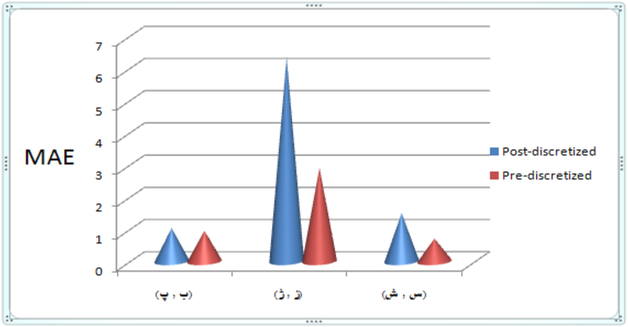

Fig. 9. MAE comparison for inter-class

6. CONCLUSION

In this work, a novel framework with the aim to recognize Hand Written Characters in the Kurdish language is proposed and the researchers have demonstrated the impact of the process of discretization. A general experiment has been effectively carried out with the implementation of the framework that is proposed. The individual or unique features in a Hand Written Character maybe systematically symbolized with the implementation of the invariants discretization algorithm. The outcome reveals that with the implementation of the invariant discretization method, the effectiveness of the Hand Written Character recognition of Kurdish characters is greatly enhanced with the organization to acquire enhanced effectiveness paralleled to pre-discretized data. In future work another experiment may be carried out on some different characters to enhance the overall effectiveness of the model.

REFERENCES

[1] Guo, X. T., Christian, V. G. and Alex, C. K.(2010). Individuality of Alphabet Knowledge in Online Writer Identification**.** *IJDAR* Springer Berlin / Heidelberg. 1433-2833.

[2] Muzaffar, B. and Jurgen, K.,(2009). Person Authentication with RDTW based on Handwritten PIN and Signature with a Novel Biometric Smart Pen Device, IEEE
Workshop. 63-68.

[3] Srihari, N. S. and Ball, R. G., (2009). Semi-Supervised Learning for Handwriting
Recognition. *ICDAR*, 26-30.

[4] 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.

[5] Behzad, H. and Mohsen, M.,(2010). A text-independent Persian Writer Identification based on Feature Relation Graph (FRG). Pattern Recognition 43. 2199–2209.

[6] Bayan Omar Mohammed , (2013). HANDWRITTEN KURDISH CHARACTER RECOGNITION USING GEOMETRIC DISCERTIZATION FEATURE , Volume 4, Number 1 , January-June 2013 pp. 51-55.

[7] Anil K. J., Robert P. W. and Jianchang, D. M. (2000). Statistical Pattern Recognition: A Review, in *Proc. 4th IEEE Trans on Pattern analysis and Machine intelligence*, 22, 4-37.

[8] AK Muda, SM Shamsuddin, A Ajith,(2010). Improvement of Authorship invarianceness for individuality representation in writer identification. Neural Netw. World **3**(10), 371–387.

[9] Bayan Omar Mohammed , (2012). ' Uniqueness in Kurdish Handwriting', International Journal of Engineering & Computer Science IJECS-IJENS Vol:12 No:06 , pp:42-50.

[10] Fabrice Muhlenbach and Ricco Rakotomalala, (2005). Discretization of Continuous Attributes. In John Wang (Ed.) Encyclopedi,a of Data Warehousing and
Mining, pp. 397-402.

[11] B. O. Mohammed , S. M. Shamsuddin ,(2012). Improvement in twins handwriting identification with invariants discretization , EURASIP Journal on Advances in Signal Processing *2012*, 2012:48 doi:10.1186/1687-6180-2012-48

[12] Bayan Omar Mohammed and Siti Mariyam Shamsuddin,(2011). Feature Discretization for Individuality Representation in Twins Handwritten Identification, Journal of Computer Science 7 (7): 1080-1087, 2011, ISSN 1549-3636, Science Publications.

[13] ID. Trier and AK. Jain. (1996). Feature Extraction Methods for Character Recognition- A Survey," Pattern Recognition, vol. 29, no. 4, 641- 662.

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

[15] K. Azmi, R. Kabir and E. Badi ,(2003). Recognition Printed Letters wit Zonong Features. Iran Computer Group, vol. 1, 29-37.

[16] R.Muralidharan,C.Chandrasekar,(2011) . Object Recognition using SVM-KNN based on Geometric Moment Invariant , International Journal of Computer Trends and Technology , ISSN: 2231-2803 , pp. 215-219.