Uniqueness in Kurdish Handwriting

Bayan Omar Mohammed
Computer Department, College of Science and Technology, University of Human Development, Kurdistan Region, Iraq

Abstract-- Kurdish writing shows such an amazing fact which has been conducted through research in the domain of pattern recognition and computer vision. Later, there were a lot of different ways and methods which have been improved or implemented to progress an approach for writer identification. The thing that's very difficult to deal with is that the handwriting of a person differs from one person to another and considering the human error it is impossible for one person to write the same thing over and over again where it has to be the exact writing. However this is considered individualistic where the consistent individual features are hidden in the handwriting. In the other hand, in the alphabet of the western countries like English, Arabic, German, Korean and Kurdish have structural characteristics, but this kind of structural characteristics is considered a bit complex in Kurdish handwriting and it is a very large set of characters. For that reason this study going to concentrate on the Kurdish Handwriting to extracting macro features for the Ward shape of the WI. The elicit of the macro features are being done thoroughly where it is important to check and investigate granulary to legitimize the presence of solo features; thus the idea of implementing the Authorship Invarianceess. Where the Authorship can be stated as a specific amount of conservation solo features neglecting the transformation of the hand writing. Where the Authorship Invarianceess will be enhanced by decreasing the similarity error for intra-class [stated by the same writer]. With the increment of the similarity error for inter-class [stated by different writer]. The solo performance is injected to the system by injecting different issues for the solo feature into standard performance or individual feature this is being achieved by take help of Invariant Discretization. Where the advantage of the Invariant Discretization is that it have the ability to enhance the identification level for each of the handwriting where our main concentration is being on the Kurdish handwriting.

Index Term-- Pattern recognition, unique representation, authorship invarianceeness, real writer, Kurdish handwriting, discretization.

1. INTRODUCTION
Pattern recognition is essential in different kind of engineering and scientific discipline to use recognition in order to differentiate between them this type of recognition is called Pattern recognition. Pattern recognition is used in computer vision, artificial intelligence and biology. Where in pattern recognition hand writing is recognized as one of the famous and various areas more specific in writes Identification (WI). Where it is well known that handwriting to be individualistic so, when writing solo the outcome depends on the hypothesis putting in mind that solo or individual have handwriting (1-6, 23-24). This kind of information gave a big push to the researchers to look for a way to find a solution to WI. But it is still stumbling block for the human capability to observe and recognize the styles of the handwriting. The study implies that what the science stated about WI converge on the extraction of the features and classification task. Execrated features from the main task for the writer appear in different representations which degrade the classification performance. However, additional steps need to be performed in order to have a better representation of input prior to the classification task. This study will identify such additional process that can transform the different representations, which are resulted from the feature extraction task, into a better representation of individual features.

2. INDIVIDUALITY OF KURDISH HANDWRITING
For any person handwriting way the individual nature can be obtained plus the hypothesis in [12-14,23] shows that for every person the handwriting is considered harmonious and consistent. The Fig. below shows that the handwriting for 3 different writers. Fig. 1 showing for the same shape and same characters and figure 2 shows the same thing but different character. Comparing the result still for the same writer there is a bit of difference between the handwritten but once there is different writer involved in the matter the difference is too much. This phenomena famed as Individuality of Handwriting. This phenomena are identified through the variances where is could be seen as one writer (intra - class) with a smaller value of a feature of the same writer were in case of different writer it will be with a higher value (inter-class) [15-17]. Individual features are considered good and acceptable if the features have the lowest similarity error for (intra-class) and highest similarity error for (inter-class) [15]. So, it is very necessary to get individual feature from the handwriting to fulfill the requirement in order to identify the handwritten authorship. The definition & discussion of this idea and concept is shown in [18] where it is considered as authorship invarianceeness.
3. **Unique representation.**

The classifier is considering the input of the process of identification to be features therefore any feature can’t satisfy the purpose of the process. A reliable feature has acquired to specify and good identification. Applying the feature which has been extracted from the process straight an direct is considered to be a collective way of the classifier. These kind of features do not represent individual features of the writer that because the writer is represented by different features. This we will head away to a big variance between the features of the intra-class (same writer) and small variance in the inter-class (different writers). Thus to develop the invariance of the authorship, taking help of another process which can be applied using of the feature before classification process. This study will carry out the discretization technique from [15] on samples of Kurdish handwriting. The technique is meant to reduce the intra-class variance of the features while increasing the inter-class variance. Fig. 3 are showing study overview where it requires the performance of the additional procedure for the identification of the Kurdish handwritten in order to get better.

**Fig. 1.** Same character by different writers

**Fig. 2.** Different character by different writers

**Fig. 3.** Framework of discretization for WI for a Kurdish handwriting
4. Feature Extraction

Macro-features which represent the global characteristic of the habit of writing and the style with its individual can be extracted and captured from an entire document [6,11,23-24]. The purpose of this Macro-feature is to identify the writer in Kurdish handwritten. 13 macro-features with the 11 initial features in the study [6, 13]. The 11 features include the entropy of grey values, the binarization threshold, number of black pixels, number of interior contours, number of exterior contours, contour slope components consisting of number of horizontal, number of positive, number of vertical and number of negative, the average height as well as the average slant. In this experiment 6 features has been used which are the entropy of the grey values, threshold binarization, black pixel number, the average height, the average slant and the average stroke width. Why Marco-feature in this study. Because of the global characteristics which been apprehended by the feature where this show the individuality of the writer especially in the writing style and habit[14]. Elaborate description of the macro feature algorithm are provided [11-13].

5. Discretization

The difficulty that we are facing is in the training instances. Where the instant set with the characteristic of distinct, descriptive feature can be categorized in classes. In the process of discretization, changing the continuous features from the discrete with a definite number of the intervals. The range between lower an upper to present the boundary. Knowing that there are different ways to state the continuous features. There is some check point it has to be verified before the starting: 1.define the number of intervals for each discrete portion where the number in general is taken randomly. 2. Decide the boundaries for the intervals where there are different methods to discretization such as Equal Information Gain, Maximum Entropy and Equal Interval Width. Different method [15]. Invariant Discretization method, proved to be effective in being accurate and more precise. The Invariants Discretization method is a supervised method. the Discretization method is starting by searching for the suitable intervals to present the information of the writer. After looking for the suitable intervals the lower and upper boundaries will be decided. Where it is essential for the number of the intervals of an image should be like the number of the feature vectors. The uniqueness of the individual can be calculated corresponding to each writer and the information saved help the to achieve simplicity of the task classification. This discretization process confirmed to accurate especially in the nonlinear presentation [19] and with the help of a interval set, where the interpretation could be done with ease by the human [20]. Also limiting the amount of data will help the composting process to be faster [21,22]. According to the authors of [15], use of post-discretized data provided higher level of identification compared to using pre-discretized data. The result of the study showed that through the application of the discretization method on the proposed integrated Moment Invariant, higher accuracy can be achieved.

5.1 Discretization method

A specific number in interval plus value representation we can compute the features in the discretization process. Discretized feature vector is coming from the representation value, ‘generalized unique features for each individual feature which we can get. The feature generalization can show the hidden feature within the individual feature in the style of the writer. The maximum and minimum range of every writer data can be split into intervals known as ‘cuts’ of equal size to get an interval. The number of feature vector columns of the feature which been extracted locate the number of intervals. The example below emphasizes six features vectors columns acquired by using the macro - feature technique. Where the six intervals are given an upper and a lower approximation and one representation value represents each interval. In the supervised discretization, The value is calculated using the class writer. An invariant feature vector which falls into an interval will have the interval’s representation Therefore the writes which have familiar invariant feature vectors shall have familiar intervals of their classes. Where it is important to know that the information and the chartarstic of the writer are not being affected by the Discretization algorithm. The algorithm only will give the feature of the invariant vector which originally was extracted from the feature extraction process in a standard representation. Fig. 4 have the algorithm of the discretization. The requirement of the Invariant discretization of the writer class information of the discretization process. The range of intervals to be calculated in the discretization of the invariant lines which uses minimum (v<sub>min</sub>) invariant feature vector and the maximum (v<sub>max</sub>) the feature of invariant vector (mult) of the writer. Minimum line of the writer is starting point (v<sub>min</sub>) invariant feature vector maximum ends (v<sub>max</sub>) the invariant of the feature vector. Where the interval is considered to the average of the invariant discretization line this as we divide by the invariant number of the vector column.

The calculation of the interval’s width (wd ) is as follows:

\[ wd = \frac{(v_{max} - v_{min})}{f} \quad (1) \]

where :

- \( v_{min} \): minimum value of invariant feature vector for a writer.
- \( v_{max} \): maximum value of invariant feature vector for a writer.
- \( f \): the number of invariant feature vector column.
Where there is a cut point at the interval of the invariant discretization which can be the width. The invariant feature vector in an interval will have the interval’s representation value. The representation value \((rv)\) of the interval of each one is considered to be the average of the intervals where its calculation is \(rv = (iv_{\text{max}} - iv_{\text{min}})/2\). Where the intervals from 1 to 5 are to present the representation value of the invariant feature vector within \(if \ v \geq iv_{\text{min}}\) and \(if \ v < iv_{\text{max}}\) while vector of the invariant feature at the range \(if \ v \geq iv_{\text{min}}\) and \(if \ v < iv_{\text{max}}\) is to be considered at the last interval category.

Where the value of the representation known as a discretized feature vector, is a one of kind feature in an individual handwriting. Fig. 5 and 6 show the transformation of the invariant feature vector into discretized feature vector for pre-and post-discretized data, respectively. It can be seen that the discretization algorithm provides discretized feature vector that shown clear illustration of an individual’s unique features, even in Kurdish handwriting.

![Fig. 4. Invariant discretization line [15].](image1)

![Fig. 5. Example of pre-discretized Kurdish datasets.](image2)
6. Simulation Result

There are 2 experiments in this study: the experiment on the authorship invariance of the Kurdish handwriting sample and the estimation accuracy of identification of Kurdish handwriting. The 1st experiment was running to assure that the discretization technique can enhance the variance of the intra-class (same writer) and inter-class (different writer) features. Where the 2nd experiment was running to decide the discretization to improve the performance of the writer identification of the Kurdish handwriting using Fuzzy logic. The data which have been used is a group of 300 data samples taken from different 20 writers from Sulaimania University, Iraq.

6.1 Authorship invariance of Kurdish handwriting

Using Mean Absolute Error (MAE) function, the authorship invariance could be obtained. In fig. 7 example of MAE calculation. There are 15 images for each writer of handwriting samples. The 1 to 6 features taken away to represent a character. The invariance of the character plus the image reference (the 1st image) acquired by MAE [15]. Where the indication of small error means the image is closed or similar as the image reference. The MAE average can be calculated as:

\[
MAN = \frac{1}{n} \sum_{i=1}^{n} |(x_i - \bar{y_i})| 
\]  

... (2)
where \( n \) is the considered as the images number; \( x_i \) is the image at this time; \( r_i \) is the reference image or the measure location; \( f \) is the features number; \( i \) is the image feature column. Where the calculation for the authorship invarianceness for post- and pre- discretized feature vectors would be acquired by analyzing the intra-class and inter-class of the value of MAE. where the outcome will imply that using post-discretized feature vector will enhance the authorship invarianceness compared as we use the prediscretized feature vector as intra-class MAE. Taking part in the post – discretized feature vector is considered to a bit less and the inter-class MAE is considered to be more of the pre-discretized feature vector. Where low MAE value for intra-class shows that the feature at a single writer is familiar during the same at high value of MAE the inter-class shows that the Kurdish handwriting feature with change from one to another. Therefore the hypothesis are assured to be corrected and the discretization process can be enhance the authorship invarianceness at the standard representation of the individuals where it is considered to one of kind of the feature so that it helps to identify the writer in Kurdish handwriting.

The fig. 8 and 9 have a comparison of the authorship invarianceness for the macro-feature technique with post- and pre-discretized data, as the fig. shows (8 & 9 respectively) the unique individual feature where the Kurdish handwritten is considered to be unrivaled in the uniquenes. MAE value intra-class (same writer) less than the MAE value of the inter-class (different writer), it fulfills the concept which says that are individualities traits in the Kurdish hand writing. Implying the vector post-discretized feature. The feature of the individual can be even better to be compared using the vector of prediscretized feature vector. Post-discretized data must have a less MAE than pre-discretized data which being considered for the intra-class (the same writer), where the postdiscretized data can give more MAE value if we compare it to the pre-discretized data for the inter-class (different writer). Fig. 10 implies that using post-discretized data will enhance the MAE value if it was compared to the inter-class.
Fig. 8. Authorship invariancess comparison for intra-class

Fig. 9. Authorship invariancess comparison for inter-class
6.2. Identification performance evaluation with Fuzzy logic classifier
Using Fuzzy logic at the kind of Kurdish handwriting in order to get the same result for this study. Fuzzy logic classifies between distance within the writer's handwriting at the same time tries to minimize the errors of misclassification. Fuzzy logic was known at 1965 [25-27], by Lotfi A. Zadeh. Its basic idea is a multivalued logic where it allows the values of the intermediate to be defined between the conventional evaluation as true/ false, yes/no, high/low, etc. Notions which can be quick to be formulated mathematically then continue by computers so that it can use more human like way[28].

Fuzzy logic can control and change the system especially with complex or industrial processes. Also to deal with an expert system and application like the classification of SAR data.

3 experiments have been examined with different number of training data also testing data. Where the 1st one used 70% training and 30% testing data from a combination of pre-discretized and post-discretized datasets. The 2nd experiment used 60% training data and 40% testing data. Fuzzy logic got the process of training by using a matrix of classification, and the overall accuracy was calculated for each the training and testing data. The result for experiment 1 & 2 summarized in table I.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Accuracy (%)</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60% Training Data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40% Testing Data</td>
<td></td>
</tr>
<tr>
<td>Fuzzy Logic</td>
<td>73.33</td>
<td>Pre Dis</td>
</tr>
<tr>
<td></td>
<td>88.33</td>
<td>Post Dis</td>
</tr>
</tbody>
</table>

The interesting thing is that post-discretized data can give an overall identification rate with an average accuracy (%) of above 80.0%. pre-discretized data give lower identification rate less that 70.0%. This assures us that with better identification and also a higher level of accuracy can be acquired by using post-discretized datasets.

7. CONCLUSION
From the review, we acknowledged that, classification phase is essential for identification task besides pre-processing and feature extraction. This phase normally required higher computation expenses and time consuming when a huge data is used.
Each of the handwriting is unique[18,23,24]. Thus, to deal with the distinctiveness of the features, a method needs a large training data samples to adapt and learn about the writer's handwritten text [29]. For this intention, some of the method consume time and cost during computation process. Another issues regarding to misclassification. Sometimes, abnormal written characters misguide the system to classify into inappropriate class. Or in other words, any letter of a class can be treated as a member of another class. Therefore, to overcome these circumstances, further studies need to be carried out to discover the best classifier in order to reduce the computation cost and maintain the efficiency of the classification phases.

After seeing the result of MAE in section 6.1 the Authorship invariancess of the Kurdish handwriting was enhanced and better using the feature of post-discretized vector for inter-class (different writer) and intra-class (same writer) that when comparing using feature pre-discretized of the vector. This fulfilling the Individuality of handwriting of the Kurdish Identification. Putting in mind that the requirement of intra-class MAE value should be less the the inter-class MAE value. The discretization process which has been provided post-discretized vector feature to illustrate make clear the individuality of each writer. It proves that the concept of Individuality of Handwriting in WI where every writer is free to use his own style writing with differs from one writer to another even in Kurdish handwriting. The standard representation of the individual features of each consists of small intra-class variance and large inter-class variance when compared to the invariant feature vectors originally extracted through the features extraction process. As proven in Sections 6.2, this contributes to the higher accuracy of identification for each individual's handwriting. Therefore, it can be concluded that through the analysis of authorship invariancess, the application of the discretization technique should be further explored in the domain of Kurdish Handwriting Identification.

REFERENCES


