Journal of Computer Science 7 (7): 1080-1087, 2011
ISSN 1549-3636

Feature Discretization for Individuality Representation in Twins Handwritten Identification

Bayan Omar Mohammed and Siti Mariyam Shamsuddin
Faculty of Computer Science and Information Systems,
University Technology Malaysia, Malaysia

Abstract: Problem statement: The study on twins is an important form of study in the forensic and biometrics field as twins share similar genetic traits. Handwriting is one of the common types of forensic evidence. Differentiating the similarities of writing of a pair of twins is critical in establishing the reliability of handwriting identification. Writing style can be used as biometric features in authenticating individual uniqueness where these unique features can be used to identify the writer, including between a pair of twins. Existing works in Writer Identification concentrate on feature extraction and the classification task in order to identify authorship. The high similarity in a pair of twins' handwriting may degrade classification performance. There should be some standards to represent these unique features before entering into the classification task which is with the use of discretization technique. **Approach:** We proposed a new framework for writer identification in terms of identifying twins' handwriting and showed the effect of discretization process on handwriting samples of a pair of twins in order to obtain individual identification. **Results:** An experiment has been done at the Sulaimania University in Iraq with fourteen pairs of identical twins where each twin provides 4 samples of handwriting for the purpose of data collecting. These samples were implemented in this research making a comparison between the new proposed framework and classic framework. **Conclusion:** Our experimental results showed that with new framework identification of handwriting of a pair of twins can be improved through the discretization process.

Key words: Pattern recognition, twins' handwriting, individuality representation, discretization technique, writer identification, macro-feature, real writer, unique code, psychological traits, human's knowledge

INTRODUCTION

Pattern recognition serves as a vital part of various engineering and scientific fields such as computer vision, biology and artificial intelligence. Handwriting analysis is an essential branch of the pattern recognition field as it has an important purpose in the courtroom and forensic document investigations (Sargur *et al*., 2008; Tan *et al*., 2010; Srihari *et al*., 2007), Signature identification (Li and Tan, 2009) as well as Iris recognition (Chowhan and Shinde, 2008). A person's handwriting is usually affected by many personal elements such as self training history plus physiological and psychological state and nature which makes distinguishing the handwriting of a pair of twins a form of study with utmost importance.

The Twins Handwriting Identification is a quite popular area of research in pattern recognition and computer vision fields as it, in some situations, provide the only means of discovering the real writer of a written text out of a group of people (Plamondon and Lorette, 1989; Srihari, 2010).

Previous studies done on biometric identification of twins such as the discriminability between the fingerprints of a pair of twins (Jain *et al*., 2002), DNA analysis (Rubucki *et al*., 2001), computational discriminability analysis on the fingerprints of twins (Liu and Sargur, 2009), show of coefficient values in individual sets as a form of unique code for a person's face (Rycchilk *et al*., 2009), prove that there are physiological traits in nature which do not change over the years. Handwriting, however, is more associated with a person's attitude and behavioral factors rather than the aspects of psychological traits which provide researchers the real motive behind the studies on handwriting (Sargur *et al*., 2008).

Corresponding Author: Bayan Omar Mohammed, Faculty of Computer Science and Information Systems, University Technology Malaysia, Malaysia

Among the obstruction mentioned, distinguishing twins' handwriting is considered one of the first of the obstacles. As the uniqueness of an individual's criteria in their handwriting has been noticed ages ago, many techniques have been established over the years, depending on human's knowledge and proficiency to sort and compare people's handwriting (Sargur *et al*., 2008).

The ability of distinguishing a pair twins' handwriting is considered an efficient mean of personality identification via each twin's own handwriting. The identification of twins' handwriting, however, is found to be much more complex compared to that of non-twins as the resemblance of the characteristics in the writing manners of twins causes huge similarities in the features of a pair of twins' handwriting forming a purpose for the identification operation. The phase is divided into two stages; the individual features analysis and the identification and capture of similar features. The two functions are then computerized and executed accordingly via the classical method of pattern recognition in order to get results both rapidly and accurately.

Pattern recognition applications are usually executed through feature extraction and classification or learning scheme (Li and Tan, 2009; Liu *et al*., 2003; 2004). The most critical and highly prioritized process in pattern recognition is capturing and selecting the desired main features. This is especially important in the case of the identification of Twins' Handwriting. The two main problems in Writer Identification (WI), Thus: first is to find the means of acquiring the main features out of different handwriting styles or very closely similar handwriting styles as a way of discovering the real writer (Xu *et al*., 2008; Bensefia *et al*., 2005; Schlapbach and Bunke, 2004; Yu *et al*., 2004; Srihari *et al*., 2002; Shen *et al*., 2002) and how to obtain the meaningful features when comparing a pair of twins' handwriting. The second problem is categorizing the features selected from the different handwriting styles and the twins' handwriting styles into the proper classes where the features belong to.

Previous studies have developed new approaches or techniques for better feature extraction and to proof the concept of individuality in handwriting. However, from the literature, it is found that most of the studies focused on how to extract the individual features between a pair of twins and not on illustrating the individual characteristics of handwriting between the twins for systematic representation.

The focal point of this study is to suggest new framework and implement the invariant discretization process of features in order to represent the individual features of writers between a pair of twins' handwriting and significantly illustrate the related features in a systematic way to provide easier classification and obtain better identification result.

MATERIALS AND METHODS

Individuality of handwriting: Handwriting has long been considered one of the means of presenting a person's individualistic nature and the writer's individuality rests on the hypothesis that each individual has consistent handwriting (Srihari *et al*., 2007; 2002; 2001). The relation of the characters, words and the shape or style of writing is very similar between a pair of twin. However, there are still unique features for each twin. These unique features can be generalized as the individual's handwriting even though there can be high similarity in a pair of twins' handwriting. Figure 1 shows the example of the handwriting samples of two pairs of identical twins and the similarity between them.

Individuality representation: Good features acting as input to a classifier are important in order to obtain good performance in the process of identification. Extracted features usually perform the classification task directly in order to identify a writer. These features do not portray the individual features of a writer between a pair of twins because the handwriting of twin has very closely similar features which lead to small variance in the handwriting between a pair of twins. Another process is needed to in order to improve the authorship invarianceness. This study will adopt the Invariant Discretization technique based on the previous work done in (Muda *et al*., 2008) to be implemented on the twins' handwriting.

This process will help in increasing the variance between the features in the handwriting of a pair of twins. An overview of a new framework which is needed as an additional procedure prior to the classification task in order to improve the performance of the identification process of twin's handwriting.

The traditional framework is shown in Fig. 2 while the new framework is shown in Fig. 3.

Feature extraction: Macro-feature captures the global characteristic of writer's individual writing habit and style. They are extracted from the entire document (Sargur *et al*., 2008; Srihari, 2010), is exploited in this work on Twins handwriting. Totally there are thirteen macro feature including the initial eleven features reported in (Srihari *et al*., 2002; Srihari, 2010).

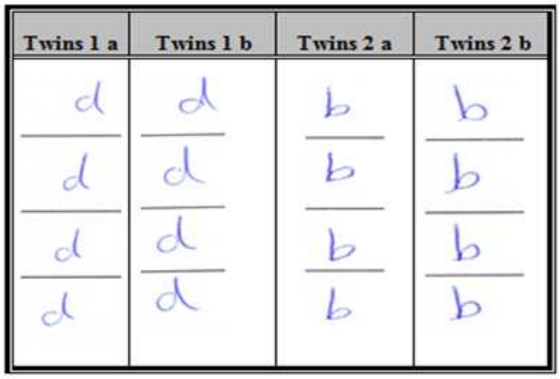

Fig. 1: Twins handwriting

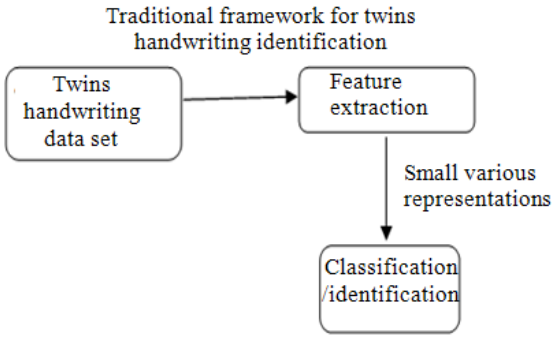

Fig. 2: Traditional framework

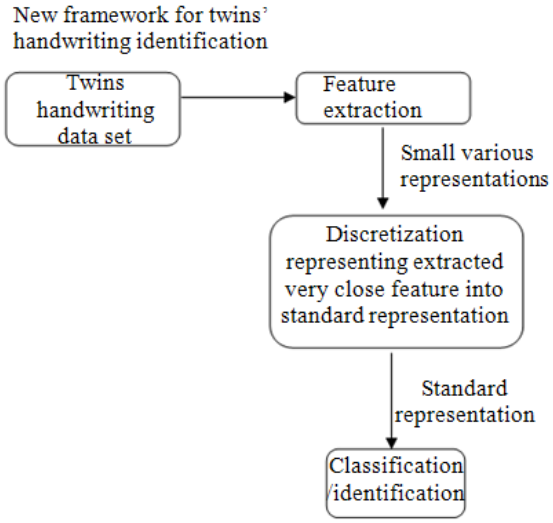

Fig. 3: New framework

The initial eleven features are: entropy of grey values, 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), average height and average slant. Eight features use in our experiment it is (entropy of grey values, binarization threshold, number of black pixels, number of interior contours, number of exterior contours, average height , average slant and average stroke width). We have chosen Macro-Features because features that capture the global characteristics of the writer's individual writing habit and style can be regarded to be macro-features (Srihari *et al*., 2007). More details on the procedure of the macro-feature algorithm can be found in (Sargur *et al*., 2008; Srihari *et al*., 2001; 2002).

Discretization: Training instances are usually the focus for classification problem. The set of training instances are usually categorized into classes with certain distinct features describing them. Through the process of discretization, discrete partition with certain number of intervals is formed when the continuous features are transformed. The range of each interval is represented by the boundaries, both lower and upper.

However, there are many ways for continuous features to be represented. This leads to the need for certain important points where firstly, the number of the intervals for the discrete partition needs to be decided. The intervals are usually selected randomly by the users. The second phase is where the boundaries of the intervals need to be determined. There are many several known methods for discretization such as Equal Information Gain (EIG), Maximum Entropy (ME) and Equal Interval Width (EIW). The recently proposed Invariants Discretization has managed to successfully provide higher rates of identification (Muda *et al*., 2008). Invariants Discretization is a supervised method where the process starts with a search for the appropriate intervals representing the information about the writer. Each interval has a set of boundaries, upper and lower. The number of the intervals for each image should also be the same as the number of feature vectors. Computation is done for each writer where according to each writer, their individual uniqueness can be preserved and the classification task can be easier done. The illustration for the feature vectors grouped with the interval will be similar to the process for the interval. The discretization process provides several benefits including non-linear representation (Agre and Peev, 2002), through the set of intervals, easier interpretation by human can be done (Liu *et al*., 2002) and with the reduced amount of data, the process of computation can be done faster and higher accuracy can be achieved (Pongaksorn *et al*., 2009; Hwang and Li, 2002). Through the study done in (Muda *et al*., 2008), it was proven that the use of discretized data can

provide better classification compared to the use of Un-discretized data. The results of the study showed a significant increase in the accuracy of the identification when the discretization is implemented on the proposed integrated Moment Invariant.

Issues regarding supervised and unsupervised discretization were discussed by Agre and Peev (2002). Two supervised method for discretization, entropy-based discretization and MVDM-based discretization, were enhanced and managed to successfully increase the accuracy of the classification process. Another study by Mehta *et al*. (2004) proposed a correlating preserving discretization which is a form of unsupervised method. The proposed algorithm was used on multivariate dataset where its efficiency proved sufficient with the prediction of missing values.

Invariant discretization process: The importance of Invariant Discretization in this work is to achieve more accurate classification of the writings of a pair of twins. With the given information about the classes of each image of the handwritings, the discretization algorithm can be applied and appropriate set of cuts representing the writer's data can be found. The minimum to maximum range of the data is divided with the size of the interval which then gives each interval or cut its lower and upper approximation. The number of the feature vectors for each image defines the number of intervals or cuts. With this, the number of applied invariant vectors in a moment invariant function can be kept to its original amount. A single representation value is defined to represent each interval or cut ensuring that their corresponding feature vectors will be similarly represented. The algorithm for the discretization process is as shown below.

Algorithm of discretization

```
For each writer {
  Min = min feature; Max = max feature;
  No_bin = no_feature_invariant;
  Interval = (Max – Min)/ No_bin;

  For each bin {
   Find lower and upper value of  interval;
   RepValue = (upper –lower)/2;
  }
  For (1 to no_feature_invariant){
   For each bin {
    If (feature in range of interval)
     Dis_Feature = RepValue;
    }
  }
}
```

Based on the writers' classes, the intervals or cuts and the values representing each interval are calculated. The Writer Identification domain set this concept where each twin in every pair has his or her own writing style and individuality which ensures the preservation of the individual's unique characteristics. The features can be illustrated much clearer and the characteristics of the features can be maintained through the process of discretization. Therefore, in order to match the process to the concept, the calculation of the intervals and representation values is done for each writer's class.

Feature invariant vector are transformed into discretize feature vector as shown in the examples in Fig. 4 and 7.

2.2	2.873	1.36	2.124	0.696	1.8	1.265	4.936	13b
2.3	2.648	1.62	2.361	0.698	1.8	1.390	9.324	13b
2.6	2.765	2.63	2.957	0.700	2.2	1.531	4.882	13b
2.4	2.089	1.55	2.248	0.688	0.9	1.812	9.544	13b
2.5	2.082	1.33	2.526	0.784	4.7	1.109	3.041	1b
2.5	2.867	1.35	2.275	0.780	2.6	1.625	3.933	1b
2.6	2.239	1.15	2.084	0.794	2.6	1.500	3.619	1b
2.9	2.468	1.72	2.398	0.784	2.6	1.734	4.234	1b
2.0	2.968	1.53	2.127	0.680	2.3	1.906	4.484	13a
2.5	2.531	2.01	2.495	0.666	1.8	1.234	1.765	13a
2.4	2.809	2.19	2.564	0.684	1.8	1.515	1.904	13a
2.1	2.831	2.32	2.524	0.686	2.3	1.421	1.915	13a
2.8	2.673	2.50	3.138	0.780	3.0	1.937	7.336	4a
2.7	2.849	1.74	2.480	0.796	3.4	1.875	7.924	4a
2.6	2.790	1.78	2.505	0.800	1.7	1.015	7.395	4a
2.6	2.488	1.84	2.588	0.796	1.8	1.031	7.244	4a
2.7	2.256	1.36	2.415	0.774	2.3	1.453	4.628	1a
2.8	2.654	1.89	2.779	0.770	2.1	1.093	1.327	1a
2.7	2.247	1.66	2.844	0.749	1.8	1.625	5.123	1a
2.9	2.204	2.03	3.149	0.741	2.6	1.281	1.602	1a
2.4	2.736	1.28	2.549	0.745	1.3	1.140	7.368	4b
2.3	2.243	1.32	2.247	0.766	1.7	1.500	7.621	4b
2.4	2.364	1.32	2.197	0.772	3.4	1.484	7.182	4b
2.4	2.110	1.31	2.406	0.764	3.6	1.078	7.055	4b
2.3	2.845	1.14	2.339	0.784	4.9	1.593	5.422	10b
2.3	2.899	0.85	1.769	0.803	2.1	3.484	5.449	10b
2.4	2.462	1.19	2.245	0.803	3.2	4.640	5.731	10b
2.3	2.621	1.07	2.150	0.811	1.7	4.125	5.810	10b
2.5	2.657	1.59	2.550	0.792	1.8	1.187	5.328	10a
2.6	2.238	1.26	2.168	0.780	1.3	1.296	5.619	10a
2.5	2.014	1.67	2.379	0.809	3.4	1.593	5.007	10a
2.5	2.740	1.37	2.467	0.794	1.7	1.218	5.370	10a
2.6	2.876	1.93	2.795	0.727	2.3	6.812	5.938	14b
2.6	2.610	1.94	2.877	0.711	2.2	6.937	3.805	14b
2.6	2.452	2.58	3.248	0.698	3.1	6.343	3.726	14b
2.7	2.660	2.55	3.166	0.711	2.2	4.171	3.330	14b
2.1	2.058	1.52	2.547	0.815	2.2	4.546	5.529	9a
2.2	3.876	1.26	2.214	0.843	1.8	4.484	4.438	9a
2.2	2.728	1.68	2.428	0.843	1.9	6.046	4.864	9a
2.0	2.910	1.99	2.472	0.843	3.7	8.015	5.455	9a
2.2	2.484	1.28	2.451	0.792	2.2	1.703	5.742	6b
2.3	2.822	1.57	2.446	0.803	1.3	5.531	1.411	6b
2.3	2.469	2.05	2.834	0.792	4.3	4.953	5.734	6b
2.3	2.734	1.62	2.579	0.802	1.4	4.953	4.867	6b
2.0	2.765	2.16	2.998	0.794	2.7	3.781	5.882	9b
2.2	2.783	1.51	2.544	0.792	1.9	3.140	5.891	9b
2.3	2.196	1.92	2.548	0.794	2.7	1.109	5.098	9b
2.0	2.447	2.06	2.571	0.807	3.2	4.421	5.723	9b
2.4	2.142	1.99	2.654	0.707	1.4	3.859	1.071	11b
2.4	2.672	1.73	2.528	0.713	2.3	3.859	5.836	11b
2.5	2.511	2.01	2.727	0.709	2.2	5.390	1.775	11b
2.4	2.860	2.36	3.107	0.703	2.7	5.328	1.430	11b
2.4	2.425	1.07	2.107	0.815	1.3	1.453	5.212	6a
1.8	2.748	1.57	2.449	0.807	1.8	4.343	5.874	6a
1.8	2.215	1.57	2.465	0.815	2.2	4.546	1.107	6a
1.8	2.426	1.57	2.364	0.819	3.6	5.328	5.213	6a
2.3	2.867	2.28	2.702	0.680	1.8	3.906	5.934	11a
2.3	2.448	2.12	2.407	0.702	1.9	5.906	5.224	11a
2.4	2.119	1.99	2.601	0.694	1.3	5.812	1.559	11a
2.4	2.473	2.39	2.908	0.692	1.4	3.109	1.236	11a

Fig. 4: Invariant feature vector data for pair of twins

Discretization for twins 1_a

d_max = 5.1230 d_min = 0.7410

Lower and Upper value for BIN for twins: 1_a

Bin 0: lower: 0.7410 Upper: 1.2888 rep value for bin 0:1.0149
Bin 1: lower: 1.2888 Upper: 1.8365 rep value for bin 1:1.5626
Bin 2: lower: 1.8365 Upper: 2.3843 rep value for bin 2:2.1104
Bin 3: lower: 2.3843 Upper: 2.9320 rep value for bin 3:2.6581
Bin 4: lower: 2.9320 Upper: 3.4798 rep value for bin 4:3.2059
Bin 5: lower: 3.4798 Upper: 4.0275 rep value for bin 5:3.7536
Bin 6: lower: 4.0275 Upper: 4.5753 rep value for bin 6:4.3014
Bin 7: lower: 4.5753 Upper: 5.1230 rep value for bin 7:4.8491

Discretization data of twins1_a

2.6581	2.1104	1.5626	2.6581	1.0149	2.1104	1.5626	4.8491
2.6581	2.6581	2.1104	2.6581	1.0149	2.1104	1.0149	1.5626
2.6581	2.1104	1.5626	2.6581	1.0149	1.5626	2.1104	4.8491
2.6581	2.1104	1.5626	3.2059	1.0149	2.6581	1.0149	1.5626

Fig. 5: Example of discretization process for twins 1_a

Discretization for twins 1_b

d_max = 4.7000 d_min = 0.7800

Lower and Upper value for BIN for twins 1_b:

Bin 0: lower: 0.7800 Upper: 1.2700 rep value for bin 0:1.0250
Bin 1: lower: 1.2700 Upper: 1.7600 rep value for bin 1:1.5150
Bin 2: lower: 1.7600 Upper: 2.2500 rep value for bin 2:2.0050
Bin 3: lower: 2.2500 Upper: 2.7400 rep value for bin 3:2.4950
Bin 4: lower: 2.7400 Upper: 3.2300 rep value for bin 4:2.9850
Bin 5: lower: 3.2300 Upper: 3.7200 rep value for bin 5:3.4750
Bin 6: lower: 4.7200 Upper: 4.2100 rep value for bin 6:3.9650
Bin 7: lower: 4.2100 Upper: 4.7000 rep value for bin 7:4.4550

Discretization data of twins1_b

2.4950	2.0050	1.5150	2.4950	1.0250	4.4550	1.0250	2.9850
2.4950	2.9850	1.5150	2.4950	1.0250	2.4950	1.5150	3.9650
2.4950	2.0050	1.0250	2.0050	1.0250	2.4950	1.5150	3.4750
2.9850	2.4950	1.5150	2.4950	1.0250	2.4950	1.5150	4.4550

Fig. 6: Example of discretization process for twins 1_b

Figure 4 shows the data for the handwriting of a pair of twins before the discretization process. The eight columns represent the extracted feature vectors while the column at the end shows the class label for the writer. A row of eight invariant vectors represent a single image for one writer.

Figure 4 shows the continuation of the discretization process done on the data in Fig. 5-6. The figure shows an example of how the data of pair a, b of twins 1 is discretized. The discretized feature data is shown in Fig. 7.

The data in Fig. 7 is the discretized feature data which shows that data representation illustrates the characteristics of each twin.

1.8591	2.8136	1.3819	2.3364	0.9046	2.3364	1.8591	4.2454	13a
2.3364	2.3364	1.8591	2.3364	0.9046	1.8591	1.3819	1.8591	13a
2.3364	2.8136	2.3364	2.3364	0.9046	1.8591	1.3819	1.8591	13a
2.3364	2.8136	2.3364	2.3364	0.9046	2.3364	1.3819	1.8591	13a
2.3485	2.3485	1.2415	2.3485	1.2415	2.3485	1.2415	4.5625	13b
2.3485	2.3485	1.2415	2.3485	1.2415	2.3485	1.2415	8.9905	13b
2.3485	2.3485	2.3485	3.4555	1.2415	2.3485	1.2415	4.5625	13b
2.3485	2.3485	1.2415	2.3485	1.2415	1.2415	2.3485	8.9905	13b
2.6581	2.1104	1.5626	2.6581	1.0149	2.1104	1.5626	4.8491	1a
2.6581	2.6581	2.1104	2.6581	1.0149	2.1104	1.0149	1.5626	1a
2.6581	2.1104	1.5626	2.6581	1.0149	1.5626	1.5626	4.8491	1a
2.6581	2.1104	2.1104	3.2059	1.0149	2.6581	1.0149	1.5626	1a
2.4950	2.0050	1.5150	2.4950	1.0250	4.4550	1.0250	2.9850	1b
2.4950	2.9850	1.5150	2.4950	1.0250	2.4950	1.5150	3.9650	1b
2.4950	2.0050	1.0250	2.0050	1.0250	2.4950	1.5150	3.4750	1b
2.9850	2.4950	1.5150	2.4950	1.0250	2.4950	1.5150	4.4550	1b
3.0125	3.0125	2.1195	3.0125	1.2265	3.0125	2.1195	7.4775	4a
3.0125	3.0125	2.1195	2.1195	1.2265	3.0125	2.1195	7.4775	4a
3.0125	3.0125	2.1195	2.1195	1.2265	2.1195	1.2265	7.4775	4a
3.0125	2.1195	2.1195	3.0125	1.2265	2.1195	1.2265	7.4775	4a
2.0343	2.8937	1.1748	2.8937	1.1748	1.1748	1.1748	7.1912	4b
2.0343	2.0343	1.1748	2.0343	1.1748	2.0343	1.1748	7.1912	4b
2.0343	2.0343	1.1748	2.0343	1.1748	3.7533	1.1748	7.1912	4b
2.0343	2.0343	1.1748	2.0343	1.1748	3.7533	1.1748	7.1912	4b
2.2922	2.8971	1.6873	2.2922	1.0824	1.6873	1.0824	5.3166	10a
2.8971	2.2922	1.0824	2.2922	1.0824	1.0824	1.0824	5.3166	10a
2.2922	2.2922	1.6873	2.2922	1.0824	3.5019	1.6873	4.7117	10a
2.2922	2.8971	1.0824	2.2922	1.0824	1.6873	1.0824	5.3166	10a
2.3546	2.9829	1.0981	2.3546	1.0981	4.8676	1.7264	5.4959	10b
2.3546	2.9829	1.0981	1.7264	1.0981	2.3546	3.6111	5.4959	10b
2.3546	2.3546	1.0981	2.3546	1.0981	2.9829	4.8676	5.4959	10b
2.3546	2.3546	1.0981	2.3546	1.0981	1.7264	4.2394	5.4959	10b
2.3219	2.9786	2.3219	2.9786	1.0084	1.6651	3.6354	5.6056	11a
2.3219	2.3219	2.3219	2.3219	1.0084	1.6651	5.6056	4.9489	11a
2.3219	2.3219	1.6651	2.3219	1.0084	1.0084	5.6056	1.6651	11a
2.3219	2.3219	2.3219	2.9786	1.0084	1.6651	2.9786	1.0084	11a
2.3071	2.3071	2.3071	2.9487	1.0238	1.6654	3.5903	1.0238	11b
2.3071	2.9487	1.6654	2.3071	1.0238	2.3071	3.5903	5.5152	11b
2.3071	2.3071	2.3071	2.9487	1.0238	2.3071	5.5152	1.6654	11b
2.3071	2.9487	2.3071	2.9487	1.0238	2.9487	5.5152	1.6654	11b
2.3904	2.3904	1.1237	2.3904	1.1237	1.1237	1.7571	4.9239	6a
1.7571	3.0238	1.7571	2.3904	1.1237	1.7571	4.2906	5.5573	6a
1.7571	2.3904	1.7571	2.3904	1.1237	2.3904	4.2906	1.1237	6a
1.7571	2.3904	1.7571	2.3904	1.1237	3.6572	5.5573	4.9239	6a
2.3389	2.3389	1.1014	2.3389	1.1014	2.3389	1.7201	5.4326	6b
2.3389	2.9576	1.7201	2.3389	1.1014	1.1014	5.4326	1.7201	6b
2.3389	2.3389	2.3389	2.9576	1.1014	4.1951	4.8139	5.4326	6b
2.3389	2.9576	1.7201	2.3389	1.1014	1.1014	4.8139	4.8139	6b
2.5494	2.5494	1.8056	2.5494	1.0619	1.8056	6.2681	4.0369	14a
2.5494	2.5494	1.8056	2.5494	1.0619	2.5494	6.2681	4.0369	14a
2.5494	2.5494	1.8056	2.5494	1.0619	2.5494	6.2681	5.5244	14a
2.5494	2.5494	1.8056	2.5494	1.0619	1.8056	6.2681	5.5244	14a
2.6477	2.6477	1.8678	2.6477	1.0879	2.6477	6.5471	5.7672	14b
2.6477	2.6477	1.8678	2.6477	1.0879	1.8678	6.5471	3.4276	14b
2.6477	2.6477	2.6477	3.4276	1.0879	3.4276	6.5471	3.4276	14b
2.6477	2.6477	2.6477	3.4276	1.0879	1.8678	4.2074	3.4276	14b
2.1650	2.1650	1.2650	2.1650	1.2650	2.1650	4.8650	5.7650	9a
2.1650	3.9650	1.2650	2.1650	1.2650	2.1650	4.8650	4.8650	9a
2.1650	3.0650	1.2650	2.1650	1.2650	2.1650	5.7650	4.8650	9a
2.1650	3.0650	2.1650	2.1650	1.2650	3.9650	7.5650	5.7650	9a
1.7481	3.0228	2.3854	3.0228	1.1107	2.3854	3.6602	5.5723	9b
2.3854	3.0228	1.7481	2.3854	1.1107	1.7481	3.0228	5.5723	9b
2.3854	2.3854	1.7481	2.3854	1.1107	2.3854	1.1107	4.9349	9b
1.7481	2.3854	1.7481	2.3854	1.1107	3.0228	4.2976	5.5723	9b

Fig. 7: Example of descritized feature data

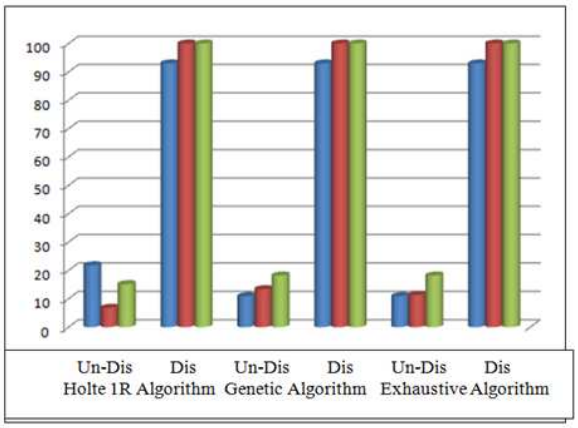

Fig. 8: Visualization of divergence level between Un-discretized and discretized twin's datasets for all data training and testing

Table 1: Comparisons of identification rates with different training and testing datasets

Rosetta built-in methods on reductions	Holte 1R algorithm		Genetic algorithm		Exhaustive algorithm	
Datasets	Un-Dis	Dis	Un-Dis	Dis	Un-Dis	Dis
Accuracy (%) 50% training data 50% testing data	21.81	92.98	10.90	92.98	10.90	92.98
Accuracy (%) 60% training data 40% testing data	6.81	100	13.36	100	11.36	100
Accuracy (%) 70% training data 30% testing data	15.15	100	18.18	100	18.18	100
Average accuracy (%)	14.59	97.66	14.14	97.66	13.48	97.66

This correctly represents the concept of individuality for every pair of twins as stated in the WI domain. The analysis of the identification performance is then done through the identification process with the discretized feature data.

RESULTS

The new framework with twin's handwriting identification successfully done. The identification of the handwriting of a pair of twins can be improved through the discretization process Fig. 8. An experiment is conducted to compare the accuracy of the identification for both discretized and undiscretized data.

The percentage of the accuracy is shown in Table 1. The extraction of the feature is done with the use of macro-features algorithm (Sargur *et al*., 2008; Srihari *et al*., 2001; 2002). Techniques provided by the Rosetta toolkit including Holte 1R classification, Genetic algorithm and Exhausive algorithm are used on the discretized and undiscretized data for the purpose of identification (Ohrn and Komorowski, 1997). The extracted features are the undiscretized data while the extracted features which have gone through the discretization process make up the discretized data. The experiment is done with the use of the samples of handwriting from fourteen pairs of twins where each twin provides four samples.

Table 1 summarizes the experimental results after running with 78 training data and 33 testing data consisting of 70% training data and 30% testing data, 67 training data and 44 testing data with 60% of which are training data and the 40% are testing data and 56 training data and 55 testing data with 50% of which are training data and 50% are testing data. The below-obtained result shows that the overall identification rate (Average Accuracy (%)) with discretized data is very good.

The overall identification accuracy from each training and testing datasets achieved is calculated through confusion matrix in (Ohrn and Komorowski, 1997).

DISCUSSION

In this study, we have introduced a new framework for twin's handwriting identification which is successfully done with discretization algorithm. The experiment for macro-features algorithm techniques for extracting the features and all tested classifiers shows that the use of discretized data produced more accurate results. The process of discretization can systematically represent the features of the data. This helps to illustrate the individuality of each twins' handwriting in discretized data more clearly. This study focuses on proving that the identification process using discretized data increases the performance of indentifying the real writer between a pair of twins' compared to the use of undiscretized data.

CONCLUSION

In this study, a new framework for the purpose of identifying twins' handwriting is proposed and we have show the effects of the discretization process on the handwriting samples of a pair of twins. An experiment has been successfully conducted with the use of the proposed framework. The individual features in the twins' handwriting can be systematically represented with the use of the invariants discretization algorithm. The extracted discrete features are used in the discretization process to granularly mine the authorship of the writer. The authorship identification can be done easier with the reduced amount of similarity errors. The results reveal that with the use of the invariant discretzation technique, the accuracy of the twins' handwriting identification is significantly improved with the overall classification get good accuracy compared to undiscretized data.

REFERENCES

Agre, G. and S. Peev, 2002. On supervised and unsupervised discretization. Cybernetics Inform. Technol., 2: 43-57. http://www.cit.iit.bas.bg/CIT_02/v2-2/43-57.pdf

Bensefia, A., T. Paquet and L. Heutte, 2005. A writer identification and verification system. Patt. Recogn. Lett., 26: 2080-2092. DOI: 10.1016/J.PATREC.2005.03.024

Chowhan, S.S. and G.N. Shinde, 2008. Iris biometrics recognition application in security management. Proceedings of the Congress Image and Signal Processing, May 27-30, Sanya, Hainan, pp: 661-665. DOI: 10.1109/CISP.2008.766

Hwang, G.J. and F. Li, 2002. A Dynamic Method for Discretization of Continuous Attributes. roceedings of the Third International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'02), Springer-Verlag, Berlin Heidelberg, London, UK., pp: 506-511. ISBN: 3-540-44025-9

Jain, A.K., S. Prabhakar and S. Pankanti, 2002. On the similarity of identical twin finger prints. Patt. Recogn., 35: 2653-2663. DOI: 10.1016/S0031-3203(01)00218-7

Li, B. and T. Tan, 2009. Online text-independent writer identification based on temporal sequence and shape codes. Proceeding of the 10th International Conference Document Analysis and Recognition, July 26-29, Barcelona, pp: 931-935. DOI: 10.1109/ICDAR.2009.26

Liu, C.-L., K. Nakashima, H. Sako and H. Fujisawa, 2003. Handwritten digit recognition: Benchmarking of state-of-the-art techniques. Patt. Recogn., 36: 2271-2285. DOI: 10.1016/S0031-3203(03)00085-2

Liu, C.-L., K. Nakashima, H. Sako and H. Fujisawa, 2004. Handwritten digit recognition: Investigation of normalization and feature extraction techniques. Patt. Recogn., 37: 265-279. DOI: 10.1016/S0031-3203(03)00224-3

Liu, H., F. Hussain, C.L. Tan, M. Dash, 2002. Discretization: An enabling technique. Data Min. Knowl. Discovery, 6: 393-423. DOI: 10.1023/A:1016304305535

Liu, Y. and N.S. Sargur, 2009. A computational discriminability analysis on twins fingerprints. Proceedings of the 3rd International Workshop on Computational Forensics (IWCF'09), Springer-Verlage Berlin Heidelberg, pp: 43-54. DOI: 10.1007/978-3-642-03521-0_5

Mehta, S., S. Parthasarathy and H. Yang, 2004. Correlation preserving discretization. Proceedings of the 4th International Conference on Data Mining, Nov. 1-4, Brighton, United Kingdom, pp: 479-482. http://doi.ieeecomputersociety.org/10.1109/ICDM.2004.10007

Muda, A.K., S.M. Shamsuddin and M. Darus, 2008. Invariants Discretization for individuality representation in handwriting authorship. Proceeding of the 2nd International Workshop on Computational Forensic (IWCF'08), Springer Verlag, Berlin, Heidelberg, pp: 218-228. ISBN: 978-3-540-85302-2

Ohrn, A. and J. Komorowski, 1997. ROSETTA: A rough set toolkit for analysis of data. Proceedings of the 3rd International Joint Conference on Information Sciences, Durham, NC, pp: 403-407. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.7135

Plamondon, R. and G. Lorette, 1989. Automatic signature verification and writer identification state of art. Patt. Recogn., 22: 107-131. DOI: 10.1016/0031-3203(89)90059-9

Pongaksorn, P., T. Rakthanmanon and K. Waiyamai, 2009. DCR: Discretization using class information to reduce number of intervals. http://conferences.telecom-bretagne.eu/data/qimie09/pongaksorn_etal_QIMIE_2009.pdf

Rubucki, R.J., B.J. McCue, K.J. Duffy, K.L. Shepard, S.J. Shepherd and J.L. Wisecarver, 2001. Natural DNA mixtures generated in fraternal twins in utero. J. Forensic Sci., 46: 120-125. PMID: 11210896

Rycchilk, M., W. Stankiewicz, M. Moezynski, 2009. Method of numerical analysis of similarity and differences of face shape of twins. Proceedings of the 13th International Conference on Biomedical Engineering (ICBE'09), pp: 1854-1857. DOI: 10.1007/978-3-540-92841-6_460

Sargur, S. H. Chen, S. Harish and S. Vivek, 2008. On the disriminability of the handwriting of twins. J. Forensic Sci., 53: 430-446. http://www.cedar.buffalo.edu/~srihari/papers/TR-04-07.pdf

Schlapbach, A. and H. Bunke, 2004. Off-line handwriting identification using HMM based recognizers. 17th International Conference on Pattern Recognition, Aug. 23-26, Cambridge, UK., pp: 654-658. http://doi.ieeecomputersociety.org/10.1109/ICPR.2004.1334343

Shen, C., X.-G. Ruan and T.-L. Mao, 2002. Writer identification using Gabor wavelet. Proceedings of the 4th World Congress on Intelligent Control and Automation (WCICA'02), IEEE, pp: 2061-2064. DOI: 10.1109/WCICA.2002.1021447

Srihari, S.N., 2010. Computational methods for handwritten questioned document examination. Ph.D. U.S. Department of Justice National Institute of Justice United States, pp: 46. http://www.ncjrs.gov/app/publications/Abstract.aspx?id=254838

Srihari, S.N., C. Huang, H. Srinivasan and V.A. Shah, 2007. Biometric and Forensic Aspects of Digital Document Processing. In: Digital Document Processing Advances in Pattern Recognition, Chaudhuri, B.B. (Ed.). Springer, London, pp: 379-405. http://www.mendeley.com/research/biometric-forensic-aspects-digital-document-processing/

Srihari, S.N., S.-H. Cha, H. Arora and S. Lee, 2001. Individuality of handwriting: A validation study. Proceeding of the 6th IAPR International Conference on Document Analysis and Recognition (ICDAR'01), Seattle, WA , USA., pp: 106-109. DOI: 10.1109/ICDAR.2001.953764

Srihari, S.N., S.-H. Cha, H. Arora and S. Lee, 2002. Individuality of handwriting. J. Forensic Sci., 47: 1-17. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.3775&rep=rep1&type=pdf

Tan, G.X., C. Viard-Gaudin and A.C. Kot, 2010. Individuality of Alphabet Knowledge in Online writer Identification. Proceedings of the 10th International Conference Document Analysis and Recognition, July 26-29, Springer Berlin Heidelberg, Barcelona, pp: 56-60. DOI: 10.1109/ICDAR.2009.165

Xu, D.-Y., Z.-W. Shang, Y.-Y. Tang and B. Fang, 2008. Handwriting-based writer identification with complex wavelet transform. Proceedings of the International Conference Wavelet Analysis Pattern Recognition, Aug. 30-31, IEEE, Hong Kong, pp: 597-601. DOI: 10.1109/ICWAPR.2008.4635849

Yu, K., Y. Wang and T. Tan, 2004. Writer identification using dynamic features. Biometric Authentication. Lecture Notes Comput. Sci., 3072: 512-518. DOI: 10.1007/978-3-540-25948-0_70