WRITER IDENTIFICATION BASED ON HYPER SAUSAGE NEURON

Samsuryadi, Siti Mariyam Shamsuddin and Norbahiah Ahmad
Universiti Teknologi Malaysia, Malaysia, syamsuryadi@unsri.ac.id
\{mariyam, bahiah\}@utm.my

ABSTRACT. This paper proposes biomimetic pattern recognition (BPR) based on hyper sausage neuron (HSN) and applies it in writer identification. HSN is used to cover the training set. HSN’s coverage can be seen as a topological product of a one-dimensional line segment and an n-dimensional supersphere. The feature extraction is moment invariants such as united moment invariants (UMI) and aspect united moment invariants (AUMI). The experiments result show that AUMI-HSN method is more effective than UMI-HSN method for identifying the authorship of handwriting.

Keywords: biomimetic pattern recognition, hyper sausage neuron, writer identification, united moment invariants, aspect united moment invariants

INTRODUCTION

Research on handwriting analysis based on the identification of the author's point of view in the last ten years experienced a significant development, particularly in forensic applications. A writer identification system aims to search a document legal ownership of a person against a large database with a sample of the author's handwriting recognition (Bulacu & Schomaker, 2007). Special image-making is done based on the features captured from each individual's handwriting. The final decision made by forensic experts to determine the identity of the author sample in question.

One of the problems of identification for purposes of the authors often appear in the court of justice in determining whether a conclusion about the authenticity of the document. This also applies in some institutions that analyze the text of former writers, and identification of various authors who took part in the preparation of the manuscript. The significant results from recent years in the field of handwriting recognition makes it possible to bring this significant answers to specific problems.

At this time, many researchers have used statistical decision model in identify the writer from the handwriting samples. Pattern classification used to determine the pattern without using some previous knowledge of the relationship between the samples in the same class. This differs from the human function.

Human being recognizes things individually by finding the commonalities between things in the same class. This is done by assuming that the sample points of the same class in the feature space would be continuous and recognizable characters. Hence, recognition of a certain class of objects is important, the analysis and cognition of the “shapes” of the infinite point sets constituted by all the objects in feature space. This concept is called biomimetic pattern recognition (BPR) by Wang Shoujue (Shoujue, 2003). BPR concept is incorporated
into writer identification for identifying authorship of handwriting (Samsuryadi & Shamsuddin, 2010).

This paper focuses on hyper sausage neuron (HSN) for writer identification. Firstly, some handwritings are extracted through united moment invariant (UMI) (Yinan, et.al, 2003) and aspect united moment invariant (AUMI) techniques (). Secondly, HSN classifier is used to identify the features obtained at the first step. The experiments of writer identification is implemented to demonstrate learning ability and the correct rate of AUMI-HSN and UMI-HSN methods.

WRITE IDENTIFICATION BASED ON HSN

Biomimetic Pattern Recognition (BPR)

In the real world, every one finds one by one similarity between things in the same class. If there are two samples belong to the same class, the differences between them should gradually change. So there must be a sequence of gradual changes between the two samples. Principle of continuity between homologous samples in feature space is called the principle of homology-continuity (PHC) (Shoujue & Xingtao, 2004). PHC can be described in mathematical formulas: suppose that point set \( A \) includes all samples in the same class \( A \) in feature space. If \( x, y \in A \) and \( \varepsilon > 0 \) are given, there must be set \( B \):

\[
B = \{x_1 = x, \ldots, x_{n-1}, x_n = y | \rho(x_i, x_{i+1}) < \varepsilon, \forall i \in [1, n-1], i \in N \} \subset A
\]  

(1)

It is a kind of prior knowledge of sample distribution in the BPR to improve the cognitive ability, then BPR intends to find the optimal covering of samples in the same class. The basic step of BPR is to analyze the relation between training samples of the same class in the feature space, which is made possible through the PHC of sample distribution (Jiang, at. al., 2009).

Cover Neuron

HSN is as the basic covering unit of the training set. HSN’s coverage in high dimensional space, which constructs a sausage like shape in feature space for covering the distribution area of the sampling points in the same class, (Shoujue & Xingtao, 2004). The HSN covering can be seen as a topological product of a one-dimensional line segment and an two-dimensional supersphere (Xu & Wu, 2010).

Cover process

Let \( A = \{A_1, A_2, \ldots, A_n\} \), is the samples points of the training set and one sample denoted \( A_i = (a_{i1}, a_{i2}, \ldots, a_{il}) \), where \( i = 1, 2, \ldots, n \) and \( l \) is dimension of the feature space or number of features.

The construction steps of HSN for writer identification are as follows:

**Step 1.** Calculate the Euclid distance every two points in the \( A \), find two points with the shortest distance, denoted \( B_{11} \) and \( B_{12} \). \( L_i \) is segment line \( B_{11}B_{12} \). HSN covers \( B_{11} \) and \( B_{12} \) is denoted as \( H_1 \), and it coverage is \( C_1 \):

\[
C_1 = \{X | \rho(X, L_i) \leq k_1, X \in R^n\} \quad (2)
\]

\[
L_i = \{Y | Y = \alpha B_{11} + (1 - \alpha) B_{12}, \alpha \in [0,1]\}
\]

(3)

where \( \rho(X, L_i) \) is the distance between the point \( X \) and the covering unit \( L_i \).
Step 2. Let \( U_1 = S - \{B_{11}, B_{12}\} \). Find point in \( U_1 \) is the nearest to \( B_{12} \), denoted as \( B_{13} \) and make the second segment line \( B_{12}B_{13} \), denoted as \( L_2 \). HSN covers \( B_{12} \) and \( B_{13} \) is denoted as \( H_2 \), and it coverage is \( C_2 : \)

\[
C_2 = \{X | \rho(X, L_2) \leq k \}, X \in \mathbb{R}^n \}
\]

(4)

\[
L_2 = \{Y | Y = aB_{12} + (1 - \alpha)B_{13}, \alpha \in [0,1]\}
\]

(5)

Step i. Delete remaining points which are included in \( C_1, C_2,..., C_{i-1} \). Find point \( B_{i(i+1)} \) in the remaining points, which is nearest to \( B_i \), denoted line segment \( B_iB_{i(i+1)} \), is as \( L_i \). HSN covers \( B_i \) and \( B_{i(i+1)} \) is denoted as \( H_i \), and it coverage is \( C_i : \)

\[
C_i = \{X | \rho(X, L_i) \leq k \}, X \in \mathbb{R}^n \}
\]

(5)

\[
L_i = \{Y | Y = aB_i + (1 - \alpha)B_{i(i+1)}, \alpha \in [0,1]\}
\]

(6)

The above algorithm is terminated, if all the points in \( A \) have been covered. Finally we have (n-1) HSNs, and the covering area of training samples in this case is the union set of the areas by these neurons:

\[
C = \bigcup_{j=1}^{n-1} C_j
\]

(7)

In this study, we adopted \( k = \beta D_y \), where \( D_y \) is the distance between \( A_j, A_j \) (Xu & Wu, 2010). \( \beta \) is in the range of [0.30, 0.75].

Identifying Algorithm

Calculate the distance \( \rho_i \) between sample point \( A \) for identifying and the union \( C_i \) of class \( i (i = 1, 2, ..., q) \) and \( \rho_i \) was defined as formula (8).

\[
\rho_i = \min_{1 \leq j \leq M_i} D_{ij}
\]

(8)

where \( D_{ij} \) was the minimum distance from \( A \) to the complex geometrical body \( C_j \) \(( j = 1, 2, ..., M_i )\) of union \( C_i \).

Calculated each \( \rho_i \) for \( A \). Finally the testing sample \( A \) would be classified to the class which corresponding to the least \( \rho_i \) namely,

\[
r = \arg \min_{1 \leq i \leq q} \rho_i
\]

(9)

RESULT AND DISCUSSION

In this paper, the handwriting data are obtained from IAM database. We choose 10 persons with 10 words were selected and each word was made for 10 times (all 1000 samples). We use two feature extraction methods such as united moment invariants (UMI) and aspect united
moment invariants (AUMI) to show that BPR is not relied on certain feature extraction method.

For each of 10 persons (writer) has 20 training samples (4 words x 5 repetition), and 25 testing samples (5 words x 5 repetition). Each training samples is used to training the neurons of BPR model for each class, thus each cover set of the 10 persons has 19 HSNs. The experiment result in percentage for beta value 0.30 as far as 0.75 can be showed in Table 1.

Table 1. Percentage Result for Each Writer and Beta Value Based on AUMI-HSN

<table>
<thead>
<tr>
<th>Writer</th>
<th>Beta value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.30</td>
</tr>
<tr>
<td>W1</td>
<td>56</td>
</tr>
<tr>
<td>W2</td>
<td>84</td>
</tr>
<tr>
<td>W3</td>
<td>32</td>
</tr>
<tr>
<td>W4</td>
<td>52</td>
</tr>
<tr>
<td>W5</td>
<td>20</td>
</tr>
<tr>
<td>W6</td>
<td>76</td>
</tr>
<tr>
<td>W7</td>
<td>64</td>
</tr>
<tr>
<td>W8</td>
<td>28</td>
</tr>
<tr>
<td>W9</td>
<td>48</td>
</tr>
<tr>
<td>W10</td>
<td>20</td>
</tr>
<tr>
<td>Average</td>
<td>48.00</td>
</tr>
</tbody>
</table>

Based on Table 1, W1 with beta value 0.30 can be identified 14 samples from 25 samples (56%), 92% (23/25) for beta value 0.40, and so on. The best average result of identifying writer from 10 writers in beta value 0.75 is 96.40%. We can see beta value has influence to identify the authorship of handwriting.

We do the same way for UMI-HSN with 10 writers, 20 training samples and 25 testing samples and the best average result in beta value 0.75 is 88.00%, detail result shows in Table 2.

Table 2. Percentage Result for Each Writer and Beta Value Based on UMI-HSN

<table>
<thead>
<tr>
<th>Writer</th>
<th>Beta value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.30</td>
</tr>
<tr>
<td>W1</td>
<td>24</td>
</tr>
<tr>
<td>W2</td>
<td>36</td>
</tr>
<tr>
<td>W3</td>
<td>8</td>
</tr>
<tr>
<td>W4</td>
<td>32</td>
</tr>
<tr>
<td>W5</td>
<td>24</td>
</tr>
<tr>
<td>W6</td>
<td>0</td>
</tr>
<tr>
<td>W7</td>
<td>40</td>
</tr>
<tr>
<td>W8</td>
<td>32</td>
</tr>
<tr>
<td>W9</td>
<td>32</td>
</tr>
<tr>
<td>W10</td>
<td>32</td>
</tr>
<tr>
<td>Average</td>
<td>26.00</td>
</tr>
</tbody>
</table>

Besides experiment above, we do the other training samples and testing samples to show the performance of the method. For instance, UMI(30,35) means 30 training samples and 35 testing samples for feature extraction, UMI and classification method, HSN (UMI-HSN) for beta values from 0.30 to 0.75. The complete result can be showed in Figure 1.
Based on Figure 1, we make difference percentage correct rate between UMI-HSN method and AUMI-HSN method for beta values 0.75 as Table 3.

Table 3. The percentage matches the identification with UMI-HSN and AUMI-HSN

<table>
<thead>
<tr>
<th>Data</th>
<th>Correct rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Samples</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>30</td>
<td>25</td>
</tr>
<tr>
<td>20</td>
<td>35</td>
</tr>
<tr>
<td>30</td>
<td>35</td>
</tr>
</tbody>
</table>

Based on Table 3, correct rate UMI-HSN method for 25 testing samples with 20 and 30 training samples has the average result decrease from 88.00 to 86.00, and 35 testing samples with 20 and 30 training samples has the average result decrease from 89.71 to 88.00. This condition is different from AUMI-HSN method, the adding number of training samples can increase the percentage correct rate result.

CONCLUSION AND FUTURE WORK

This paper proposed AUMI-HSN and UMI-HSN for identifying the authorship of handwriting. The experiments in Table 3 showed that AUMI-HSN method was better than UMI-HSN method, the correct rate UMI-HSN was around 88% and AUMI-HSN was around 96%. Future work can be conducted to further explore the moment invariants feature extraction methods and cover neurons appropriate for BPR.
ACKNOWLEDGMENT

Authors would like to thank Research Management Centre (RMC) under vote RJ13000078284F026 and Soft Computing Research Group (SCRG), Universiti Teknologi Malaysia, for the support in making this study a success.

REFERENCES


