Mathematical Handwritten Formula Recognition

Chaninthorn Amornsawaddirak  
School of Information, Computer and Communication Technology,  
Sirindhorn International Institute of Technology, Thammasat University  
Email: c.amornsawaddirak@student.siit.tu.ac.th

Cholwich Nattee  
School of Information, Computer and Communication Technology,  
Sirindhorn International Institute of Technology, Thammasat University  
Email: cholwich@siit.tu.ac.th

Nirattaya Khamsemanan  
School of Information, Computer and Communication Technology,  
Sirindhorn International Institute of Technology, Thammasat University  
Email: nirattaya@siit.tu.ac.th

Abstract—With the recent advances in pen-based computing and optical scanning technologies, we have all the necessary hardware components for interacting with computer systems via digital pens or stylus. The development of online handwriting recognition technique also allows us to efficient input text messages by handwriting. Developing a system that allows inputting mathematical expressions by handwriting is another challenging research task. This paper presents feature extraction in online handwritten recognition for mathematical symbols. This method is based on Hidden Markov Models to compare accuracy rates of feature extractions. There are 11,224 samples in this experiment that are grouped into 4 group. One group is into test group and others are into train group. The experimental result shows that average recognition rates of using four features (normalized distance to stroke edge, normalized y-coordinate, vicinity slope, and curvature slope) are higher than others.

I. INTRODUCTION

Nowadays, Information Technology is one of the most essential things in people’s daily life. Everything can be done through handheld devices. We can use them to do a lot of things such as typing the document, taking the note and so on. The input of character on devices is not only done by keyboard but also by using handwriting. According to study[1], the physical movement of the handwriting helps brains memorize and understand things more efficient than typing. Therefore, handwriting recognition plays an important role for communicating with devices using pen.

The handwriting recognition is to input data of mathematical expressions into a computer or tablet. Online handwriting recognition is a process that aims to identify the most appropriate character from handwritten strokes. It is done by extracting features from a set of handwritten strokes, then apply a machine learning technique to build a model that classifies a stroke into a class of symbols. To efficiently recognize an expression, we will also focus on analyzing the structure of mathematical expressions, so that the recognized symbols can be correctly combined into an expression.

In this paper, we focus on developing a on-line handwriting recognition technique in particular for mathematical formulas based on Hidden Makov Model to compare accuracy rates of feature extraction techniques.

II. LITERATURE REVIEW

Liwicki, 2009 [2] presents the best features from their experiment applying on a feature set in order to discover which features are significant for handwriting recognition. A Hidden Markov Model (HMM) and a bidirectional long short-term memory network (BLSTM) were used for handwriting recognition. The experiment results show that there are the cosine of the slope, the normalized y-coordinate, the density in the center of the context maps, the pen-up/down information, and the size of the curvature.

In order to solve problem of disappearance when pen up, Nakai, 2002 [3] proposes a technique based on the interpolation of pen pressure feature. There are two features related to pen pressure: the first is the pressure representing pen ups and downs in a continuous manner; the other is the interpolating pen pressure during the pen-up interval to find interpolating the pen pressure between the pen leaves tablet and next pen down in one stroke.

The Speed Normalization is a preprocessing technique which is implemented to remove the variation of writing velocity. Pastor, 2005 [4] proposes two methods for Writing Speed Normalization technique. The first is trace segmentation which it can be used to normalize the writing speed. This method is controlled by a parameter called resampling distance by using a fixed distance interval. The second method is derivatives normalization. The writing speed can be normalized by using the derivative module in each point. The module of normalized derivatives will be constant (equal to 1).

Hu, 2011 [5] presents a technique for mathematical symbols recognition by using HMM and Segmental K-means. The segmental K-means algorithm is an algorithm for estimating the hidden markov model parameters by embedding the K-means method. This experiment uses a variant of segmental K-means to get initialization of Gaussian Mixture Model' parameters of observation probability distribution. The result shows that average recognition rates of using a variant of segmental K-means are higher.

Jirakunkanok, 2012 [6] uses Conditional Random Fields (CRF) technique to classify Thai cursive handwriting. They compare the accuracy of CRF with that of HMM and the result shows that CRF outperforms HMM.

III. BACKGROUND

There are several recognition techniques for handwriting recognition. In this thesis, we focus on the hidden markov model (HMM).

A hidden Markov model (HMM) is a statistical Markov model in which is assumed to be a Markov process with
hidden states. It is a generative probabilistic model and used to compute the probability of a sequence of class variables \( y \) for an observation sequence \( x \). A state transition probability matrix depends on the previous state and observation probability distribution depends on the current state.

IV. METHODOLOGY

A. Preprocessing

1) Deleted duplication point: duplicated point is the point that has the same \((x,y)\) coordinates and it cannot give any useful information for classification. (in Fig. 1 and 2)

2) Size normalization: eliminating the variation of \( y \)-coordinate’s size by transforming the \( y \) coordinates range to be \([0,1]\) but reserve the ratio of stroke. (in Fig. 3)

It can be computed as:

\[
y' = \frac{y - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}} \tag{1}
\]

\[
x' = \left(\frac{x_{\text{max}} - x_{\text{min}}}{y_{\text{max}} - y_{\text{min}}}\right)\left(\frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}\right) \tag{2}
\]

3) Smoothing: reducing the noise information by average of each point with previous point, current point, and following point. (in Fig. 4)

4) Speed normalization: removing the influence of writing velocity by the method in [4]. (in Fig. 5)

B. Feature extraction

1) Pen up/down: this feature is a binary feature whether the stylus has contact with the tablet or not at time \( t \). (in Fig. 6)

2) Normalized distance to stroke edge: In order to add the location information to the pen-up/down feature by taking the distances to the beginning and the end of the stroke. (in Fig. 7) It can be computed as:

\[
N(s,t) = \begin{cases} 
1 - \frac{|d_e - d_b|}{l_s} & \text{for actual stroke} \\
-\left(1 - \frac{|d_e - d_b|}{l_s}\right) & \text{for interpolated stroke} 
\end{cases} \tag{3}
\]

where \( l_s \) represents the length of the stroke, \( d_b \) represents the distance between the current point and the first point of stroke, and \( d_e \) represents the distance between the current point and the last point of stroke.

3) Normalized \( y \)-coordinate: The \( y \)-coordinate’s position after size normalization.

4) Vicinity slope \( \alpha \): The vicinity slope of the current point \((x(t), y(t))\) is represented by the cosine and sine of the angle \( \alpha \). (in fig. 8) It can be computed as:
\[ \cos \alpha = \frac{\Delta x_1 \Delta y_1 + \Delta x_2 \Delta y_2}{(\sqrt{\Delta x_1^2 + \Delta x_2^2})(\sqrt{\Delta y_1^2 + \Delta y_2^2})} \] (4)

5) Curvature \( \beta \): The curvature of the current point \((x(t),y(t))\) is represented by cosine and sine of the angle \(\beta\). (in fig. 8) It can be computed as :

\[ \cos \beta = \frac{\Delta x_1 \Delta y_1 + \Delta x_2 \Delta y_2}{(\sqrt{\Delta x_1^2 + \Delta x_2^2})(\sqrt{\Delta y_1^2 + \Delta y_2^2})} \] (5)

where \(P(y,x)\) is the joint distribution probability, \(y\) is a sequences of states, \(x\) is an observation sequences, \(p(y_i|y_{i-1})\) is the transition probability distribution, \(p(x_i|y_i)\) is the observation probability distribution, and \(N\) is the number of distinct states in HMM.

HMM is defined by three parameters:

\[ \pi_i = P(y_1 = 1) \] (7)

\[ a_{ij} = P(y_t = 1|y_{t-1} = 1) \] (8)

\[ b_{ik} = P(x_t = 1|y_t = 1) \] (9)

where \(\pi_i\) is the prior probability of the initial state being state \(i\), \(a_{ij}\) is the transition probability, and \(b_{ik}\) is the observation symbol probability.

For training, we construct the HMM model which composes of initial distribution probabilities, observation probabilities and state transition probabilities. Three probabilities are used to compute the probability of a sequence of states for an observation sequence.

For testing, the probability associated with each of the HMM is calculated and the symbol whose HMM has the maximum probability is the recognition result.

V. Experiments and Results

The handwriting dataset used in this experiment was extracted from Competition on Recognition of Online Handwritten Mathematical Expressions. Table I shows all the 56 symbols which can be recognized in our system. There are 11,224 strokes that we grouped them into 4 groups. We chose one group into the test group and others were combined into the train group.

This experiment aimed to evaluate the performance of recognition on feature extraction technique. In experiment, we used combination formula for selecting feature extraction to compare the accuracy of handwriting recognition. The result of accuracy recognition is in Table II.

where

feature 1 represents pen up/down information.
feature 2 represents normalized distance to stroke edge.
feature 3 represents normalized y-coordinate.
feature 4 represents vicinity slope.
feature 5 represents curvature slope.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Group1</th>
<th>Group2</th>
<th>Group3</th>
<th>Group4</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.27%</td>
<td>2.60%</td>
<td>2.13%</td>
<td>6.40%</td>
<td>3.85%</td>
</tr>
<tr>
<td>2</td>
<td>18.96%</td>
<td>12.40%</td>
<td>15.00%</td>
<td>15.38%</td>
<td>15.44%</td>
</tr>
<tr>
<td>3</td>
<td>10.34%</td>
<td>16.83%</td>
<td>15.61%</td>
<td>15.59%</td>
<td>14.48%</td>
</tr>
<tr>
<td>4</td>
<td>16.41%</td>
<td>18.63%</td>
<td>17.85%</td>
<td>19.42%</td>
<td>18.07%</td>
</tr>
<tr>
<td>5</td>
<td>14.34%</td>
<td>12.82%</td>
<td>12.31%</td>
<td>16.79%</td>
<td>14.06%</td>
</tr>
<tr>
<td>1.2</td>
<td>17.73%</td>
<td>17.91%</td>
<td>17.88%</td>
<td>15.84%</td>
<td>17.34%</td>
</tr>
<tr>
<td>1.3</td>
<td>30.77%</td>
<td>31.64%</td>
<td>31.89%</td>
<td>32.40%</td>
<td>31.67%</td>
</tr>
<tr>
<td>1.4</td>
<td>23.62%</td>
<td>22.16%</td>
<td>26.52%</td>
<td>25.54%</td>
<td>24.46%</td>
</tr>
<tr>
<td>1.5</td>
<td>11.52%</td>
<td>24.35%</td>
<td>19.94%</td>
<td>16.69%</td>
<td>18.13%</td>
</tr>
<tr>
<td>2.3</td>
<td>37.56%</td>
<td>36.25%</td>
<td>33.88%</td>
<td>40.61%</td>
<td>37.07%</td>
</tr>
<tr>
<td>2.4</td>
<td>68.51%</td>
<td>68.03%</td>
<td>67.26%</td>
<td>66.75%</td>
<td>67.64%</td>
</tr>
<tr>
<td>2.5</td>
<td>27.89%</td>
<td>29.46%</td>
<td>28.65%</td>
<td>33.28%</td>
<td>29.82%</td>
</tr>
<tr>
<td>3.4</td>
<td>72.42%</td>
<td>73.83%</td>
<td>70.31%</td>
<td>65.16%</td>
<td>70.43%</td>
</tr>
<tr>
<td>3.5</td>
<td>37.01%</td>
<td>37.07%</td>
<td>38.96%</td>
<td>35.79%</td>
<td>37.21%</td>
</tr>
<tr>
<td>4.5</td>
<td>18.02%</td>
<td>19.98%</td>
<td>16.78%</td>
<td>18.11%</td>
<td>18.23%</td>
</tr>
<tr>
<td>1.2.3</td>
<td>38.17%</td>
<td>33.46%</td>
<td>31.39%</td>
<td>35.19%</td>
<td>34.55%</td>
</tr>
<tr>
<td>1.2.4</td>
<td>65.46%</td>
<td>64.23%</td>
<td>64.27%</td>
<td>71.35%</td>
<td>66.34%</td>
</tr>
<tr>
<td>1.2.5</td>
<td>34.04%</td>
<td>35.25%</td>
<td>34.30%</td>
<td>37.99%</td>
<td>35.39%</td>
</tr>
<tr>
<td>2.3.4</td>
<td>72.85%</td>
<td>70.47%</td>
<td>70.96%</td>
<td>73.68%</td>
<td>73.40%</td>
</tr>
<tr>
<td>2.3.5</td>
<td>58.38%</td>
<td>58.32%</td>
<td>55.53%</td>
<td>59.17%</td>
<td>57.85%</td>
</tr>
<tr>
<td>1.3.4</td>
<td>58.20%</td>
<td>60.38%</td>
<td>60.79%</td>
<td>53.73%</td>
<td>58.28%</td>
</tr>
<tr>
<td>3.4.5</td>
<td>59.53%</td>
<td>58.67%</td>
<td>58.62%</td>
<td>60.24%</td>
<td>59.27%</td>
</tr>
<tr>
<td>1.4.5</td>
<td>55.62%</td>
<td>58.74%</td>
<td>61.93%</td>
<td>60.45%</td>
<td>59.18%</td>
</tr>
<tr>
<td>2.4.5</td>
<td>56.48%</td>
<td>53.52%</td>
<td>49.80%</td>
<td>56.38%</td>
<td>54.05%</td>
</tr>
<tr>
<td>1.3.5</td>
<td>51.74%</td>
<td>56.91%</td>
<td>58.55%</td>
<td>61.62%</td>
<td>57.21%</td>
</tr>
<tr>
<td>1.2.3.4</td>
<td>89.33%</td>
<td>86.52%</td>
<td>86.99%</td>
<td>85.85%</td>
<td>87.17%</td>
</tr>
<tr>
<td>1.2.3.5</td>
<td>56.23%</td>
<td>53.34%</td>
<td>57.77%</td>
<td>57.36%</td>
<td>56.31%</td>
</tr>
<tr>
<td>1.2.4.5</td>
<td>73.47%</td>
<td>72.83%</td>
<td>73.90%</td>
<td>73.89%</td>
<td>73.52%</td>
</tr>
<tr>
<td>1.3.4.5</td>
<td>81.00%</td>
<td>84.19%</td>
<td>84.25%</td>
<td>84.01%</td>
<td>83.36%</td>
</tr>
<tr>
<td>2.3.4.5</td>
<td>89.47%</td>
<td>88.88%</td>
<td>88.69%</td>
<td>90.06%</td>
<td>89.28%</td>
</tr>
<tr>
<td>1,2,3,4,5</td>
<td>87.79%</td>
<td>89.38%</td>
<td>88.77%</td>
<td>89.21%</td>
<td>88.79%</td>
</tr>
</tbody>
</table>

Table II. The Accuracy Recognition of Each Feature Extraction

VI. CONCLUSION AND FUTURE WORK

This study proposes a recognition technique for on-line handwriting recognition based on Hidden Markov Model. The proposed system is evaluated the performance of recognition on feature extraction techniques. The result shows that average recognition rates of using four features (normalized distance to stroke edge, normalized y-coordinate, vicinity slope, and curvature slope) are the highest at 89.28%, followed by using five features at 88.79% and using four features (pen up/down, normalized distance to stroke edge, normalized y-coordinate, and vicinity slope) at 87.17%.

For our future works, we aim to recognize the mathematical symbol by using the Conditional Random Fields Model to compare with Hidden Makov Model.

ACKNOWLEDGMENT

This research is financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA) and the National Research University Project, Thailand Office of Higher Education Commission.

REFERENCES