A Hybrid Feature Extraction Method for Accuracy Improvement in “Aksara Lontara” Translation

Intan Sari Areni, Asyraful Insan Asry and Indrabayu

Department of Electrical Engineering, Universitas Hasanuddin, Indonesia
Department of Informatics, Universitas Hasanuddin, Indonesia
Research Group of Artificial Intelligence and Multimedia Processing (AIMP)

Article history
Received: 19-04-2017
Revised: 15-07-2017
Accepted: 26-07-2017

Abstract: An Optical Character Recognition (OCR) of “Aksara Lontara” has been constructed using a novel combination of feature extraction methods in this study. The ancient font of “Lontara” is then translated into Bahasa Indonesia to help non-native language to learn this language. Two powerful extraction feature methods, i.e., Modified Direction Feature (MDF) and Fourier Descriptor (FD) are stages combined to deal with two dominant phases of the Lontara font. The classification process is conducted using Support Vector Machine (SVM) as a fast and straightforward learning method deal with 23 fonts in image containing of 150×120 pixels. In this research, 50 verbs were used for training and 30 verbs for validating the system. The results show that system can reach 96% accuracy using this hybrid in extraction feature with kernel variable of $C = 3$ and $\sigma = 8$.

Keywords: Lontara script, Modified Direction Feature, Fourier Descriptor, Support Vector

Introduction

One of a great and most implemented field in pattern recognitions is Optical Character Recognition (OCR) which can transform images into a script for further processing by computers (Mori et al., 1992). Nowadays, this OCR technique is also exploited for non-Latin fonts such as Japanese kanji (Budiwati et al., 2011) and Chinese (Hao et al., 2011). Due to uniqueness in font pattern of each language, a researcher needs to find and explore the best methods from preprocessing until recognition stages.

“Aksara Lontara” is one of the ageless legacies of Indonesia, particularly in Bugis and Makassar tribes. It is written in lontara leaves and preserves until now. “Lagaligo” is one of the famous books written in lontara, fills with poetry and philosophy of two biggest kingdoms at the time i.e., Buginese and Makassarnese (Mattulada, 1991). An effort for digitalization and embedding this literacy into compact and straightforward form is vital for the new generation. This portable system will be useful not only for scholars but also for tourism purposes related to Aksara Lontara.

Several researchers have engaged in developing an OCR for aksara lontara through assimilating image processing techniques and artificial intelligence. It has been completed for specific characters and classification purposes only using Backpropagation Neural Network (Alwi and Wardoyo, 2010). Moreover, a web base of aksara lontara for translation was adapted using parsing tree and binary search methods (Nangi et al., 2013). Modified Direction Feature (MDF) was proposed for aksara Lampung recognition which is limited to stand-alone primary characters and without punctuation (Prarian et al., 2013). Another method, Fourier Descriptor (FD), was applied in detection of handwritten Aksara Batak Toba using Multilayer Perceptron algorithm neural network (Sihombing, 2013), which has been multiplied somewhat for classification stage in artificial intelligence application (Indrabayu et al., 2013).

In that paper, Learning Vector Quantization is used for classification step (Tjokorda et al., 2009). Another method, Fourier Descriptor (FD), was applied in detection of handwritten Aksara Batak Toba using Multilayer Perceptron algorithm neural network (Sihombing, 2013), which has been multiplied somewhat for classification stage in artificial intelligence application (Indrabayu et al., 2013).

Previous researches related to aksara recognition just focus on recognize one character without punctuation and MDF method was well performed on that. In this study, the authors proposed to recognize two characters of aksara Lontara that have special rules in reading and...
writing. In this study, authors proposed a hybrid feature extraction, MDF and FD methods, for a translation system of “Aksara Lontara” based on image script. This is a uniqueness of this article. FD method can detect pixels of shape character (Rajput and Horakeri., 2011) especially for punctuation shapes in Lontara word.

The translation system is built for Indonesia tourism especially in Makassar.

The rest of this paper is organized as follows; Section 2 describes methodology include discussion of the hybrid feature extraction implementation. Section 3 explained results and discussion of the case study. Section 4 describes our conclusion and some future works.

Methodology

Words in Lontara formed by two or more Lontara symbols which have special rules in reading and writing. This study focuses on the translation of the verb which is composed of two symbols of lontara. Figure 1 shows the character of Lontara script.

Image of Lontara verb is used as data input with a size of 150×120 in JPG format. Font type is BugisA.ttf with a size of 14 pt. The classification stage consists of two processes, i.e., training and testing process. The number of the used words for training and testing is 50 and 30, respectively. The training process is applied to get the best parameters of the feature values which is obtained by the extraction process of MDF and FD methods. When the training process has been carried out, the best outcome parameters is saved and observed for the further classification of testing data. The result of the testing process is associated with a database for the translation process. The proposed system design is shown in Fig. 2.

Preprocessing

Preprocessing consists of grayscaling and binarization. Grayscaling is the process of converting the pixel values of RGB images to grayscale pixel value based on the following equation:

\[ \text{Gray} = w_R R + w_G G + w_B B \]  

where, \(w_R\), \(w_G\) and \(w_B\) are the weight value of Red, Green and Blue color, respectively. According to National Television System Committee (NTSC), \(w_R\) is 0.299, \(w_G\) is 0.587 and \(w_B\) are 0.114. Figure 3 shows a sample of the RGB values for the word "Akka" at the point marked.

![Fig. 1. Character of Lontara script](image1)

![Fig. 2. Proposed system design](image2)
Table 1. The value of label directions ($d_n$)

<table>
<thead>
<tr>
<th>Direction</th>
<th>Label</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertical</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Right Diagonal</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Horizontal</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Left Diagonal</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

MDF method is combination feature method of FD (Fourier Descriptor) and FT (Fourier Transform). In FD, the feature value is determined by the direction value ($d_n$) as shown in Table 1.

After the gray scaling process, each pixel value is converted into pixel values range from 0-255. From this point, the grayscale value $G(x,y)$ of each pixel is converted into a binary value $B(x,y)$, which is called binarization process, with a threshold value ($T$) of 210 and based on the following equation:

$$B(x,y) = \begin{cases} 1 & \text{if } G(x,y) \leq T \\ 0 & \text{if } G(x,y) > T \end{cases}$$

(2)

Hybrid Feature Extraction

Modified Direction Feature (MDF) (Liu and Blumenstein, 2004) and Fourier Descriptor (FD) (Kadir & Susanto, 2013) methods are used in feature extraction stage. At the MDF, there are two feature values ($N_j$), namely direction feature value ($f_d$) and the transition function value ($f_t$) where a flow diagram of MDF is shown in Fig. 4.

Furthermore, the value of $f_t$ is calculated by the following equation:

$$f_t = 1 - \left( \frac{p}{w} \right)$$

(3)

where, $p$ is pixel position and $w$ is image width. While the value of $f_d$ is determined by the direction ($d_n$) with the following equation:

$$f_d = 0.1 \ d_n$$

(4)

The $d_n$ value is derived from Table 1 based on the direction value of object-forming pixels. Figure 5 shows the $d_n$ value example of the word "Akka". While the $f_t$ and $f_d$ values for the word of "Akka" is shown in Fig. 6.

The value of $f_t$ and $f_d$ is normalized in each direction with a normalization matrix size ($N_w$) of $5 \times 3$ and those values are combined from all sides so that the resulting vector length is 120 ($= N_w \times N_l \times N_d$).

After the features extraction results obtained by the MDF, the following process is also conducted using FD. The features extraction with FD is determined by the value example of the word "Akka". The value of $f_t$ and $f_d$ is normalized and based on the following equation:

$$U_n = \frac{1}{N} \sum_{j=0}^{N-1} s(t). \exp \left( -\frac{j2\pi t}{N} \right), n = 0, 1, 2, \ldots, N - 1$$

(5)
where, $N$ is the number of image pixels, $U_n$ is the Fourier descriptors which consist of the object pixel values with $n = 0, 1, 2, \ldots, N-1$. The FD stage is shown in Fig. 7.

The combination of features extraction concept is shown in Fig. 8. The MDF method can detect every pixel in a gray area, but not in a small contour (red punctuation). The deficiencies in the MDF can be overcome by the FD method. Thus, the combination of both methods will improve the performance of the proposed translator system.

![Fig. 5. The label directions object value of the word “Akka”](image5.png)

![Fig. 6. Feature values: (a) Transition feature value ($f_t$) and (b) Direction feature value ($f_d$)](image6.png)

![Fig. 7. Fourier descriptor stages](image7.png)

![Fig. 8. The combination of feature extraction concept](image8.png)
Classification

The resulting feature values on all training data will be combined and used as an input data in the training process with Support Vector Machine (SVM). SVM method aims to lock the best hyperplane to get the path separator from two different classes (Cheriet et al., 2007). This study uses SVM multi-classes to classify each feature values with the estimation method of one to one so that the separation function of the two classes can be determined. There are two parameters must be considered in SVM training with the Radial Basis Function (RBF) kernel, i.e., C and γ. The C parameter controls the influence of each individual support vector and the γ parameter defines how far the influence of a single training example reaches.

Furthermore, the training data is expressed as \((x_i, y_j)\), with \(i = 1, 2, 3, \ldots, M\), where \(M\) is the number of data and denoted by the model of \(x_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iq}\}^T\) that is an attribute to \(i\)-th training data. Class labeling is denoted by \(y_j \in \{-1, +1\}\), where \(j = 1, 2, 3, \ldots, P\). A class of data samples separated by hyperplane is expressed by the following equation:

\[
f(x) = wx + b
\]

(6)

where, \(w\) is weighting vector and \(b\) is a constant value.

where, \(\gamma = 1/2\sigma^2\) with \(\sigma\) is a free parameter.

The performance of translation system is expressed with an accuracy level \((A_L)\) based on the following equation:

\[
A_L = \left(\frac{N_r}{N_t}\right) \times 100\%
\]

(7)

where, \(N_r\) is the amount of correct data and \(N_t\) is the amount of all data.

Results and Discussions

In the testing process, the used testing data is 30 by varying parameters \(C\) and \(\sigma\) in kernel RBF. Figure 9 shows the accuracy of the MDF and MDF-FD for any difference parameter.

![Fig. 9. The comparison accuracy of the translation system; (a) \(C = 1\); (b) \(C = 2\); (c) \(C = 3\); (d) \(C = 4\).](image-url)
Table 2. The results of the testing data

<table>
<thead>
<tr>
<th>No</th>
<th>Word Aksara</th>
<th>Spelling</th>
<th>Meaning</th>
<th>MDF</th>
<th>MDF-FD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jamma</td>
<td>Adu</td>
<td>Adu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Bissa</td>
<td>Basuh</td>
<td>Bicara</td>
<td>Basuh</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Akka</td>
<td>Angkat</td>
<td>Angkat</td>
<td>Angkat</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Tunu</td>
<td>Bakar</td>
<td>Bakar</td>
<td>Bakar</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Dince</td>
<td>Tancap</td>
<td>Tancap</td>
<td>Tancap</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Giling</td>
<td>Bilik</td>
<td>Bilik</td>
<td>Bilik</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Mala</td>
<td>Ambil</td>
<td>Ambil</td>
<td>Ambil</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Lurang</td>
<td>Angkat</td>
<td>Angkat</td>
<td>Angkat</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Tojang</td>
<td>Ayun</td>
<td>Ayun</td>
<td>Ayun</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Era</td>
<td>Ajak</td>
<td>Ajak</td>
<td>Ajak</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Walek</td>
<td>Balas</td>
<td>Balas</td>
<td>Balas</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Jallo</td>
<td>Amuk</td>
<td>Amuk</td>
<td>Amuk</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Tahang</td>
<td>Tahan</td>
<td>Tahan</td>
<td>Tahan</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Bici</td>
<td>Bisik</td>
<td>Bisik</td>
<td>Bisik</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Bali</td>
<td>Bantu</td>
<td>Makan</td>
<td>Bantu</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Nyikka</td>
<td>Ikat</td>
<td>Makan</td>
<td>Ikat</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Pau</td>
<td>Bicara</td>
<td>Bicara</td>
<td>Bicara</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Ekbu</td>
<td>Buat</td>
<td>Buat</td>
<td>Buat</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Bampa</td>
<td>Pukul</td>
<td>Pukul</td>
<td>Pukul</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Manre</td>
<td>Makan</td>
<td>Makan</td>
<td>Makan</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Gerrak</td>
<td>Bentak</td>
<td>Bentak</td>
<td>Bentak</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Sungke</td>
<td>Buka</td>
<td>Buka</td>
<td>Buka</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Boya</td>
<td>Cari</td>
<td>Cari</td>
<td>Cari</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Nange</td>
<td>Renang</td>
<td>Renang</td>
<td>Renang</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Sare</td>
<td>Beri</td>
<td>Beri</td>
<td>Beri</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Tawa</td>
<td>Bagi</td>
<td>Bagi</td>
<td>Bagi</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Basa</td>
<td>Bicara</td>
<td>Bicara</td>
<td>Bicara</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Anang</td>
<td>Anyam</td>
<td>Angkat</td>
<td>Angkat</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Powe</td>
<td>Belah</td>
<td>Belah</td>
<td>Belah</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Tettong</td>
<td>Berdiri</td>
<td>Berdiri</td>
<td>Berdiri</td>
<td></td>
</tr>
</tbody>
</table>

In Fig. 9, the best accuracy using MDF is 86% when $C = 3$ and $\sigma = 8$, which is improved by using the combination of feature extraction MDF-FD and scaling up to 96% with the same parameter. Table 2 shows the results of each data testing.

Misclassification with MDF feature extraction occurs in the data that has punctuation on top due to the normalization value of the $f_i$ and $f_t$ will tend to ignore small fractal value. This issue will lead to mistranslation.

**Conclusion**

The process of feature extraction with a combination of MDF-FD system has been applied to the translation system on Lontara script with image-based data input. The total amount of data process is 80 comprises of 50 training data and 30 testing data. The results showed that the performance of the translation system increased 10% (from 86 to 96%) with a hybrid features extraction when $C = 3$ and $\sigma = 8$. This is due to FD method can recognize the small shape in Lontara script that is not detected by MDF. Hence, the combination of both methods can improve the system performance.

In the future research, this system will incorporate more data, not only for verbs but also nouns. Words meaning will also seek into single and two connected words. Moreover, this system will be verified by other methods, such as Wavelet and Shearlet transforms.

For tourism purposes, this study will be improved by building an Android-based application to get a portable and user-friendly translation tools.
Acknowledgement

This work was supported by WCU Grant of Universitas Hasanuddin.

Author’s Contribution

All authors equally contributed to this work.

Ethics

This article is the original contribution of the authors and is not published elsewhere. There is no ethical issue involved in this article.

References


Prarian, C., I.T. Iwan and R. Magdalena, 2013. Character recognition of Aksara lampung analysis using Modified Direction Feature (MDF) and hidden Markov Model. Faculty of Electrical Engineering, Telkom University, Bandung.

