Balinese glyph recognition with gabor filters

M W A Kesiman¹, G A Pradnyana² and I M D Maysanjaya³

¹,²,³Laboratory of Cultural Informatics, Department of Informatics, Faculty of Technology and Vocational, Universitas Pendidikan Ganesha, Jalan Udayana No. 11 Singaraja, Bali, Indonesia, 81116

¹antara.kesiman@undiksha.ac.id

Abstract. Recognizing Balinese glyphs from the Balinese script on palm leaf manuscripts is not trivial. In Balinese script, there are more than a hundred glyphs which represent basic syllables and compound syllables, and also some punctuation marks. They naturally share a strong interclass similarity between each other related to the form of their writing curves. The degraded image of textured palm leaf manuscript also offers some challenging parts in recognizing the Balinese glyph. In this paper, we investigated the use of Gabor filter bank as the feature extraction method to recognize the Balinese glyphs. By using Gabor filter, we can detect many texture variations with different orientations and frequencies. In our experiments, the published dataset of AMADI_LontarSet for glyph recognition was used. It showed a very promising result by using a single hidden layer Neural Network as the classifier. Gabor filters with Zoning method achieved a high enough recognition rate. For future works, Gabor filters will be analyzed in combination with the Histogram of Gradient, Neighborhood Pixel Weight and Kirsch Edges features.

1. Introduction
In many document image analysis (DIA) research projects, glyph recognition or character recognition is frequently performed. Naturally, the glyph/character recognition step is needed in the Optical Character Recognition (OCR) module, especially for the segmentation based text recognition methods. For the segmentation based text recognition methods [1], [2], prior to recognition process, a text line will be segmented into individual or isolated glyph/character segment. The recognition process will be performed for each glyph/character segment.

Many works on the isolated character recognition were already reported. The existence of wide variety of feature extraction methods to recognize a glyph/character helps many DIA projects to achieve their optimal performances. Nevertheless, the findings of many ancient documents or manuscripts all over the world offer a continuous challenges for glyph/character recognition methods. These ancient manuscripts contain many ancient scripts with totally a new collection of glyph/characters that need to be analysed. For example, the ancient manuscript collections found in some Asian countries or more specifically in some Southeast Asian countries. In the cases of these manuscripts, the feature extraction methods which are already widely used for Latin script recognition are not trivial to be applied and to be implemented. This is due to the fact that those ancient manuscripts with the ancient scripts, glyphs and characters have many specific characteristics in writing style, different form and formats, and oftenly with many very complex writing rules. For example the Chinese and Japanese script families [3]–[6], the Indian script families [7], [8], the Gurmukhi script [9]–[12], the Sundanese script [13], the Khmer script [14]–[16], and the Balinese script [17]–[19].
The collection of manuscripts on palm leaves from Bali, Indonesia, offers also a significant number of challenges for the DIA research projects [17]. These manuscripts were written in dried palm leaves and scratched with a small knife to write the script. The first challenge is the physical condition of the manuscripts. They are degraded and the writings are faded. The second challenge is the complexity of the Balinese script as it is a family of the alpha-syllabic script. This paper presents an investigation study on the use of Gabor filters to recognize the isolated Balinese glyph. Previous work on the isolated Balinese glyph recognition has actually been reported [18]. But, a new challenge in the word spotting system research for the Balinese manuscript collections on palm leaves leads to a new hypothesis in the use of Gabor filters for feature extraction method.

Section 2 of this paper presented a short overview about the characteristics of the Balinese glyphs in Balinese script. Gabor filters were presented in Section 3. The experimental results, the Balinese glyph dataset and the evaluation method was described in Section 4. Finally, Section 5 some conclusions and future works were listed in Section 5.

2. Balinese Glyphs of Balinese Script

In this section, a brief overview of the use of Balinese script in Balinese palm leaf manuscripts was presented. The Balinese glyph collection was also introduced.

2.1. The use of Balinese Script in Manuscripts on palm leaves

More than fifty thousand collections of manuscripts on palm leaves were found in Bali, Indonesia (Figure 1). The main collections, about six thousands palm leaf manuscript collections can be found in two principal museum owned by the Balinese regional government. But, most of the remaining collections are owned by the Balinese private family at their houses. The important values of these collections are the variety of the manuscript's content. It ranges widely from the religious contents, arts, laws, architectural aspects, traditional medical knowledge, and so many socio-cultural aspects of the ancient Balinese ways of life.

![Figure 1. Three examples of Balinese palm leaf manuscript from three different regions.](image)

The Balinese language, mixed with the old Javanese of Kawi and the Sanskrit, with the Balinese script were used in palm leaf manuscripts writing. The Balinese script is a descendant of the Brahmi script. Nowadays, the Balinese in Bali speak the Balinese language, but they do not write in Balinese script. It makes the young generation is not totally aware of the richness of their ancient cultural heritages in the palm leaf manuscript collections. They do not know how to write in Balinese script, and it makes even more difficult to read the Balinese script in the manuscript collections.

2.2. Balinese Glyphs
In Balinese script, there are more than a hundred glyphs. They consist of the consonants, the independent vowels, the dependent vowels, the conjunct form of the consonants, the digits, the punctuations, and many additional symbols for special consonants and musical symbols (Figure 2).

![Figure 2. The complete Balinese glyph composition in Balinese script.](image)

Writing in Balinese script (Figure 3), the glyphs can be written in upper position of the medial text line as the ascender glyphs and in lower position of the medial text line as the descender glyphs. Balinese script writing is considered as one of the most complex script in the world.

![Figure 3. An example of how to write “windu” in Balinese script.](image)

3. Gabor Filters

Gabor filter is a filter modulation between sinusoid and Gaussian filter [20]. Gabor filter is a texture based filter with many orientations and frequencies. A bank of Gabor filters is a collection of many Gabor filters with different orientation, wavelength, aspect ratio and bandwidth as the parameters (Figure 4). In our hypothesis, Gabor filter can provide the initial information about the existence of textures in the document. It can be used to detect the preliminary informations about text and non text
area on the palm leaf manuscript images. In the palm leaf manuscripts, the writing shows the spatial texture informations where the different writers represent the different frequencies and orientations of the textures.

![Gabor filters visualization by Prasad, V.S., & Domke, J. (Taken from [20]).](image)

Figure 4. Gabor filters visualization by Prasad, V.S., & Domke, J. (Taken from [20]).

In this work, Gabor filters will be combined with Zoning method as feature extraction method. In the preprocessing step, each greyscale image of isolated Balinese glyph segment was resized into 50 x 50 pixels. A bank of Gabor filters was finally applied. It consisted of 64 different Gabor filters. The 64 Gabor filters were generated by combining the four main parameters of Gabor filter with different parameter values such as the orientation from 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°, the values for wavelength were 4 and 8, 0.5 and the value for aspect ratio was 0.9, and the values for bandwidth were 0.5 and 1. The greyscale Gabor filtered images were then binarized by using the global binarization method of Otsu [21]. The Zoning method was finally applied to the binarized image of Gabor filtered image. It consisted of seven different zoning area such as vertical and horizontal zoning, block zoning, diagonal zoning, radial and circular zoning (Figure 5). For all zoning areas, the parameter of zone width was set to 10 pixels. The feature value in each zoning area was computed by calculating the ratio of text pixels and all pixels in zoning area. Finally, all feature values which were extracted from the binarized images of the 64 Gabor filtered images, divided with seven variations of Zoning area were then joined to output a vector of 4,992 dimensions as the final feature.

![The different Zoning methods.](image)

Figure 5. The different Zoning methods.

4. Experimental Results and Evaluation

We present in this section the dataset that we used in our experimental tests to evaluate the performance of Gabor filters in recognizing the Balinese glyphs.

4.1. Dataset

To investigate the performance of Gabor filter for the Balinese glyph recognition, the dataset of AMADI_LontarSet [17] was used. This dataset is the first published dataset for Balinese palm leaf manuscript collection and it is publicly open for scientific use. For the glyph recognition work, the dataset included 133 classes of Balinese glyph, comes from 19,383 total glyph samples. 11,710 were
selected and used as train set by random selection. The remaining 7,673 samples become the test set. Each class has different number of sample images. Some samples were easily found in this dataset, but some others were too rare (Figure 6).

$$\text{Figure 6.}$$ Samples of character level annotated patch images of Balinese script on palm leaf manuscripts.

4.2. Results

Table 1 shows the recognition rate of the isolated Balinese glyph recognition by using Gabor filters and Zoning method. It shows a very promising result by using a single hidden layer Neural Network as the classifier. Gabor filters with Zoning method achieved a high enough recognition rate.

| No | Method                      | Feature Dimension | Classifier                          | Recog. Rate (%) |
|----|-----------------------------|-------------------|-------------------------------------|-----------------|
| 1. | Gabor Filters + Zoning      | 4.992             | Neural Network                      | Recog. Rate     |
|    |                             |                   | $h = 100$ (the size of hidden layer), | Train set = 97.8822 |
|    |                             |                   | $\text{reg} = 0.001$ (the regularization strength) | Test set = 70.6243 |
|    |                             |                   | $\text{step}_\text{size} = 0.01$   |                 |
| 2. | Gabor Filters + Zoning      | 4.992             | Neural Network                      | Recog. Rate     |
|    |                             |                   | $h = 200$ (the size of hidden layer), | Train set = 97.199 |
|    |                             |                   | $\text{reg} = 0.001$ (the regularization strength) | Test set = 71.9145 |
|    |                             |                   | $\text{step}_\text{size} = 0.01$   |                 |
| 3. | Gabor Filters + Zoning      | 4.992             | Neural Network                      | Recog. Rate     |
|    |                             |                   | $h = 300$ (the size of hidden layer), | Train set = 97.088 |
|    |                             |                   | $\text{reg} = 0.001$ (the regularization strength) | Test set = 71.2889 |
|    |                             |                   | $\text{step}_\text{size} = 0.01$   |                 |

5. Conclusion and Future Works
Recognizing Balinese glyphs from the Balinese script on palm leaf manuscripts is not trivial. In this paper, we investigated the use of Gabor filter bank as the feature extraction method to recognize the Balinese glyphs. Gabor filters provide initial information about the existence of text textures in the document. By using Gabor filter, we can detect many texture variations with different orientations and frequencies. For future works, Gabor filters will be analyzed in combination with the Histogram of Gradient, Neighborhood Pixel Weight and Kirsch Edges features.

References

[1] R. Chamchong and C. C. Fung, “Character segmentation from ancient palm leaf manuscripts in Thailand,” in HIP ’11 Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, 2011, p. 140.
[2] M. Muralikrishna and D. V. R. Koti Reddy, “An OCR-character segmentation using Routing based fast replacement paths in Reach Algorithm,” 2011, pp. 1–7.
[3] Z. Jin, K. Qi, Y. Zhou, K. Chen, J. Chen, and H. Guan, “SSIFT: An Improved SIFT Descriptor for Chinese Character Recognition in Complex Images,” in 2009 International Symposium on Computer Network and Multimedia Technology, 2009, pp. 1–5.
[4] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, “Online and offline handwritten Chinese character recognition: Benchmarking on new databases,” Pattern Recognition, vol. 46, no. 1, pp. 155–162, Jan. 2013.
[5] Y. Shao, C. Wang, and B. Xiao, “Fast self-generation voting for handwritten Chinese character recognition,” International Journal on Document Analysis and Recognition (IJDAR), vol. 16, no. 4, pp. 413–424, Dec. 2013.
[6] F. Yin and C.-L. Liu, “Handwritten Chinese text line segmentation by clustering with distance metric learning,” Pattern Recognition, vol. 42, no. 12, pp. 3146–3157, Dec. 2009.
[7] U. Pal and B. B. Chaudhuri, “Indian script character recognition: a survey,” Pattern Recognition, vol. 37, no. 9, pp. 1887–1899, Sep. 2004.
[8] U. Pal, R. Jayadevan, and N. Sharma, “Handwriting Recognition in Indian Regional Scripts: A Survey of Offline Techniques,” ACM Transactions on Asian Language Information Processing, vol. 11, no. 1, pp. 1–35, Mar. 2012.
[9] G. S. Lehal and C. Singh, “A Gurmukhi script recognition system,” in Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, 2000, vol. 2, pp. 557–560.
[10] A. Aggarwal, K. Singh, and K. Singh, “Use of Gradient Technique for Extracting Features from Handwritten Gurmukhi Characters and Numerals,” Procedia Computer Science, vol. 46, pp. 1716–1723, 2015.
[11] D. Sharma and P. Jhajj, “Recognition of Isolated Handwritten Characters in Gurmukhi Script,” International Journal of Computer Applications, vol. 4, no. 8, pp. 9–17, Aug. 2010.
[12] K. Singh Siddharth, R. Dhir, and R. Rani, “Handwritten Gurmukhi Numeral Recognition using Different Feature Sets,” International Journal of Computer Applications, vol. 28, no. 2, pp. 20–24, Aug. 2011.
[13] M. Suryani, E. Paulus, S. Hadi, U. A. Darsa, and J.-C. Burie, “The Handwritten Sundanese Palm Leaf Manuscript Dataset From 15th Century,” in 14th IAPR International Conference on Document Analysis and Recognition, Kyoto, Japan, 2017.
[14] EFEO, “Khmer Manuscript - Recherche - http://khmermanuscripts.efeo.fr/,” http://khmermanuscripts.efeo.fr/. [Online]. Available: http://khmermanuscripts.efeo.fr/ [Accessed: 20-Feb-2018].
[15] D. Valy, M. Verleysen, and K. Sok, “Line Segmentation for Grayscale Text Images of Khmer Palm Leaf Manuscripts,” in 7th International Conference on Image Processing Theory, Tools and Applications (IPTA 2017), Montreal, Canada, 2017.
[16] D. Valy, M. Verleysen, S. Chhun, and J.-C. Burie, “A New Khmer Palm Leaf Manuscript Dataset for Document Analysis and Recognition – SleukRith Set,” in 4th International Workshop on Historical Document Imaging and Processing, Kyoto, Japan, 2017.
[17] M. W. A. Kesiman, J.-C. Burie, J.-M. Ogier, G. N. M. A. Wibawantara, and I. M. G. Sunarya, “AMADI_LontarSet: The First Handwritten Balinese Palm Leaf Manuscripts Dataset,” in 15th International Conference on Frontiers in Handwriting Recognition 2016, Shenzhen, China, 2016, pp. 168–172.

[18] M. W. A. Kesiman, S. Prum, J.-C. Burie, and J.-M. Ogier, “Study on Feature Extraction Methods for Character Recognition of Balinese Script on Palm Leaf Manuscript Images,” in 23rd International Conference on Pattern Recognition, Cancun, Mexico, 2016.

[19] M. W. A. Kesiman, J.-C. Burie, and J.-M. Ogier, “A Complete Scheme Of Spatially Categorized Glyph Recognition For The Transliteration Of Balinese Palm Leaf Manuscripts,” in 14th IAPR International Conference on Document Analysis and Recognition, Kyoto, Japan, 2017.

[20] V. Shiv Naga Prasad and J. Domke, “Gabor filter visualization,” University of Maryland, Technical Report, 2005.

[21] “Global image threshold using Otsu’s method - MATLAB graythresh - MathWorks France.” [Online]. Available: https://fr.mathworks.com/help/images/ref/graythresh.html?requestedDomain=true. [Accessed: 20-Feb-2018].