Indonesian Journal of Electrical Engineering and Computer Science
Vol. 4, No. 2, November 2016, pp. 479 ~ 485
DOI: 10.11591/ijeecs.v4.i2.pp479-485

Balinese Script's Character ReconstructionUsing Linear Discriminant Analysis

Made Sudarma*[1], Sri Ariyani[2], Manuh Artana[3]
[1,2]Department of Electrical and Computer Engineering, Faculty of Engineering, Udayana University
Jimbaran Campus, Bali-Indonesia, Ph/Fax: +62361703315
[3]Magister Program of Electrical and Computer Engineering, Udayana University Graduate Program
Jl. PB Sudirman Denpasar 80232, Bali-Indonesia
Corresponding author, e-mail:imasudarma@gmail.com*[1], sriariyani@unud.ac.id[2],
manuhartana@gmail.com[3]

Abstract

Balinese people have one of the civilization histories and cultural heritage are handwritten in Balinese script on palm leaves known as Balinese Papyrus (LontarAksara Bali). Until now that cultural heritage is still continuously strived its preservation along with the implementation begin to be abandoned in public life. Some of Balinese Papyrus now begins to rot and fade under influenced by age. Information technology utilization can be a tool to solve the problems faced in the preservation of the Balinese papyrus. By using digital image processing techniques, the papyrus script can be reconstructed digitally so that it can be retrieved and store the content in the digital media. Balinese papyrus reconstructed through several processes from scanning into a digital image, performing preprocessing for image quality improvement, segmenting the Balinese characters on image, doing character recognition using LDA algorithm, rearranging the result of recognition in accordance with the original content in papyrus, and translating that characters result into Latin. LDA algorithm quite successfully performs the classification associated with handwritten character recognition.

Keywords*: Balinese script, papyrus reconstruction, character recognition, lda algorithm*

1. Introduction

There are many historical and cultural heritage spread all over the world including in the form of the ancient text. Such ancient manuscripts are owned by the Balinese is written on palm leaves and written in Balinese script known as Balinese papyrus [1-2]. It became one of the traditional documentation tool that has been handed down by the ancestors of Balinese people in the past that contains everything about Balinese culture [1-4].

Papyrus (Lontar) is used as a media for writing the Balinese not durable to be stored for many years. Some of papyrus has been damaged, so it is quite difficult to recognize the Balinese scripts are contained therein [3]. Information technology utilization necessary to assist in the preservation of Balinese script on papyrus by digitizing to be an image files. At the expert level, the preservation of Balinese script can be done by performing character or handwriting recognition on the image files of the papyrus [4]. By utilizing character or handwriting recognition, we can reconstruct the original contents [5, 6] of the Balinese papyrus.One of the research about handwriting recognition had conducted by Kauleshwar P., et al [7]. In his research, mentioned that Neural Network methods typically used to perform handwriting recognition related to high noise tolerance.

Research on image processing in Balinese script on papyrus, such as papyrus image segmentation, papyrus image enhancement and character recognition/ OCR has been done by some researchers [2-4], [8-9] but has never been done to reconstruct the content of the Balinese papyrus. The experiments in this study is about how to do papyrus image reconstruction by using Linear Discriminant Analysis (LDA) to recognize the characters in the papyrus. The recognition results then reassembled by using the Unicode of Balinese script. In 2006, Balinese script has been registered in Unicode in the Range 1B00-1B7F [10].The research about transliteration of Balinese script using Unicode conducted by Imam Habibi and

Received August 19, 2016; Revised October 16, 2016; Accepted October 30, 2016

Rinaldi [9]. In the last step, the reconstruction results are translated into Latin using Balinese script romanization standard rules.

Balinese script character has a high degree of similarity between each other and some of the characters are distinguished only by one line strokes [3, 8]. LDA is one of the methods used in statistics, pattern recognition [11] in general to find a linear combination of features that characterize or separating two or more classes of objects or events [12-15]. The resulting combination can be used as a linear classifier [12]. This method has proven quite successful in doing classification associated by handwriting recognition [14]. Related to handwriting and character recognition using LDA, there are some research had done before. Kurt, Z [14] had conduct research on recognition of Ottoman alphabet using LDA. Hassan E [13] had also utilizing the SDA to make OCR (Optical Character Recognition). The latest, Yuvika Dhillon [15] have used the LDA and combined with the Neural Network to character recognition research.

2. Research Method

2.1. Lontar (Papyrus)

Lontar (Javanese: ron tal, "daun tal") is siwalan leaf or tal (Borassus flabellifer or palmyra) that are drained and used as a script and handicraft. Lontar are mostly found in Bali besides in Java, Lombok, and Sulawesi islands. Lontar in Bali is used to write about Hindu Religion and cultural heritage of the ancestors [2]. There are many kind of Lontar that are grouped by function and its use. The type of the lontar such as Lontar Yajna, Lontar Wariga, Lontar Puja, Lontar Tattwa and others shown in Figure 1.

Figure 1. Lontar Wariga Gemet

Table 1. Aksara Suara (vowels)

NO	Vowel Sign	Name	Romanized	NO	Vowel Sign	Name	Romanized
1	N/a	N/a	a	7	◌ᬾ	Taleng	e
2	◌ᬸ	Tedong	ā	8	◌ᬾᬩ	Taleng-repa	ai
3	◌ᬺ	Ulu	i	9	◌ᬾᬸ	Taleng-tedong	o
4	◌ᬻ	Ulu-sari	ī	10	◌ᬾᬩᬸ	Taleng-repa tedong	au
5	◌ᬼ	Suku	u	11	◌ᬿ	pepet	ĕ
6	◌ᬽ	Suku-ilut	ū	12	◌ᬾᬿ	Pepet tedong	ö

Table 2. Aksara Wianjana (Consonant)

NO	Regular Form	Appended Form	Romanized	NO	Regular Form	Appended Form	Romanized
1	ᬳ	◌ᬵ	ha or a	10	ᬮ	◌ᬯ	La
2	ᬦ	◌ᬧ	na	11	ᬫ	◌ᬰ	Ma
3	ᬘ	◌ᬙ	ca	12	ᬕ	◌ᬖ	Ga
4	ᬭ	◌ᬬ	ra	13	ᬩ	◌ᬨ	Ba
5	ᬓ	◌ᬔ	ka	14	ᬗ	◌ᬘ	nga
6	ᬤ	◌ᬣ	da	15	ᬧ	◌ᬦ	pa
7	ᬢ	◌ᬡ	ta	16	ᬯ	◌ᬮ	ja
8	ᬲ	◌ᬱ	sa	17	ᬬ	◌ᬭ	ya
9	ᬯ	◌ᬮ	wa	18	ᬜ	◌ᬝ	nya

2.2. Balinese Script

One of the scripts are owned by Indonesia is the Balinese script or natively known as Aksara Bali and Hanacaraka. The Balinese script is also called Abugida or alpha-syllabic. It's a segmental script that is based on consonants with mandatory vocal notation but is secondary. According On Pesamuhan Agung in 1963 determined the letters of the Balinese script, on Table 3 to 6 namely Aksara Suara (vowel) and Aksara Wianjana (consonant letters) [1]. There are 47 letters in the Balinese script, each representing a syllable with inherent vowel /a/ or/ə/ at the end of a sentence. Pure Balinese can be written with 18 consonant letters and 9 vowel letters. Each consonant has a conjunct form called gantungan which nullifies the inherent vowel of the previous syllable. Punctuation includes a comma, period, colon, as well as marks to introduce and end section of a text. Texts are written from left to right without whitespace character which text generally use Balinese Language.

Table 3. Independent Vowels

No	Form	Name	Romanized	No	Form	Name	Romanized
1		a kara	a	6		u kara tedong	ū
2		a kara tedong	ā	7		e kara	e
3		i kara	i	8		arsani a	ai
4		i kara tedong	Ī	9		o kara	o
5		u kara	u	10		o kara tedong	au

Table 4. Number Characters

NO	Glyph	NO	Glyph
0		5	
1		6	
2		7	
3		8	
4		9	

Table 5. Semi Vowel Chararcters

NO	Semi Vowel Sign	Name	Romanized	Remarks
1		Guwung	ra	Same glyph as *gantungan ra*
		Guwung macelek	rĕ	Only happened when arda suara ra + ĕ (pepet)
		Guwung macelek tedong	rö	Only happened when arda suara *ra + ŏ* (pepet tedong)
2		Suku kembung	ua	Same glyph as *gantungan wa*
3		Gantungan la	la	Same glyph and name as *gantungan la*
4		nania	ia	Same glyph as *gantungan ya*

Table 6. Aksara Swalalita Chararcters

NO	Regular Form	Appended Form	Name	Romanized
1			Na rambat	ṇa
2			Da madu	dha
3			Ta tawa	tha
4			Ta latik	ţa
5			Sa saga	ṣa
6			Sa sapa	śa
7			Ga gora	gha
8			Ba kembang	bha
9			Pa kapal	pha

2.3. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is one of the methods that is used on statistical pattern recognition to determine a linear combination of features that characterize or separating two or more classes of objects or events. LDA is one of the methods used in statistics, pattern recognition in general to find a linear combination of features that characterize or separating two or more classes of objects or events. The resulting combination can be used as a linear classifier [12]. It is simple, mathematically robust and often produces models whose accuracy is as good as more complex methods. The general LDA approach is very similar to a Principal Component Analysis (PCA) [13], [15]. LDA is based upon the concept of searching for a linear combination of variables that best separates two classes. LDA can be simplified into five main steps, as detailed below.

At the first, start with computing d-dimensional mean vector.

$$\mu_i = \begin{bmatrix} \omega_i \\ \omega_i \\ \omega_i \end{bmatrix} \quad (1)$$

Where i = 1,2,3 of the class

Now, compute the two 4x4-dimensional matrices: between class scatter matrix SB and within class scatter matrix Sw. if in the PCA is computed the average a whole images only, then in the LDA we should compute the average image contained in one class. The following formula to find between class scatter matrix.

$$SB = \sum_{i=1}^{c} N_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (2)$$

μ is the overall mean, and μ_i and N_i are the sample mean and sizes of the respective classes. Next step, compute within class scatter matrix

$$S_W = \sum_{i=1}^{c} N_i(X_k - \mu_i)(X_k - \mu_i)^T \quad (3)$$

Where X_k = image k. The third step, find the covarian matrix. Covarian matrix can be obtained from operation S_B and S_W.

$$C = S_B S_W^{-1} \quad (4)$$

Next step, check the Eigen vector and Eigen value. Quick checks that the eigenvector-eigenvalue calculation is corrects and satisfy the equation:

$$C\vartheta = \lambda\vartheta \quad (5)$$

Where $C = S_B S_W^{-1}$

ϑ=Eigen vector λ=Eigen value.

At the last step, compute the LDA feature. The following formula to find LDA feature.

$$f = \sum_{i=1}^{m} (I_i - u)^T \times V \quad (6)$$

Where I =data each pixel of image training, m = total of image training.

2.4. Purpose Scheme

The design of this research is organized into some main process among others digitizing images, preprocessing, image segmentation, feature extraction, recognition and romanization Balinese scripts to produce phrase that has meaning and can be understood. In the process of preprocessing, there are several step such as brightness normalization stage, cropping ROI (Rest of Interest) of papyrus image, normalization size, adding color space to CIELAB color space, thresholding, noise reduction, and thinning using Thinning Zhang Zuen. Segmentation process split each character in the image of the Balinese papyrus. The next process is doing features extraction of each character using LDA which will be stored and used in the classification process (recognition).

In the classification process, features extraction of each character will be recognized with training set features. Previously, we divide the images of Balinese Papyrus into two groups, namely images training set and a test images with ratio 70:30. After that, the recognition results then translated into Latin script. The conversion process or transliteration has been done with standard rules of romanization Balinese script to Latin. The final result of this system is digital Balinese script arrangement and the translation shown in Figure 2.

1. Preprocessing

Preprocessing process is a pretreatment for the Balinese papyrus images that will be the object of this research. In this research will be done with some preprocessing stage. At the first, preprocessing is done by the brightness normalization process to produce uniformity of image brightness level. Next step, cropping ROI (Rest of Interest) from the images papyrus.

Then, the process of size normalization to obtain uniformity of size. Then, perform color space conversion process to CIELAB color space which aims to produce the best color in which the background images has a high color difference the desired object [2]. Followed by a thresholding process to clean and reinforce the differences in background color with the desired object. At the last, thinning process using thining Zhang-Suen [3], [8], dilation and erosion. Thinning is done to produce scratches object with uniform thickness

2. Segmentation

Balinese script papyrus images that have been done preprocessing process, will then start the segmentation process. This segmentation process aims to separate each character Balinese script writing. The results obtained from this process is a collection images of each character Balinese script contained in the image of papyrus tested. The separation is done to get the features of each character and then will be compared resemblance to a character in the Unicode Balinese script. The segmentation process is done by using Projection Profile methods. This method is done in two processes, Horizontal Projection Profile and Vertical Projection Profile [8].

3. LDA

Linear Discriminant Analysis (LDA) in this study aims to obtain an efficient way of presenting the Balinese script character space with the training data are divided into several classes or categories of Balinese script character. The image data in the training data are divided into several classes according to the group by using this information. This method can be done with 5 easy step. First count the d-dimensional mean vector, then compute the within class and between class scatter matrices, then compute the LDA covariant matrix, and then compute the eigen vector and eigen value, at the last step specify the LDA features [12].

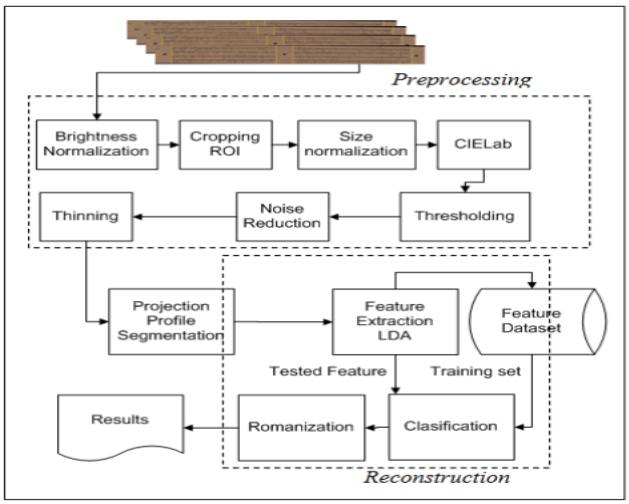

Figure 2. System Overview

4. Balinese Script Romanization

There are some rules when doing transliteration Balinese script to Latin (Romanization) [1] as detailed below.

a. There are 18 basic syllables used to write in Balinese script. Each syllable has the additional form (gantungan), which will turn off the sound/vocal syllables previously. Gantungan written under the preceding syllable.
b. Character ha often serves as a neutral character for vocals, in this case h is not transcribed. In general, ha in word-initial or vowel-medial position at the root of the word is transcribed without the h. Ha at the end of the root word followed by the suffix vowel, always transcribed by h. Occurrences h in a word also adjusted to the standard dictionary Balinese language with Latin.

c. There are 12 vowels in Balinese script which is written attached to the syllable. For stand-alone vowel in writing attached to the syllable ha. Suku and suku ilut can also be added on a gantungan or gempelan.
d. In addition to 18 basic syllables, letters bali has more syllables called aksara Swalalita, the rules are just as basic syllables.
e. Signs tedung can change the shape of the base syllables, except syllable ba, nga, ja and nya.
f. The whole basic syllable may have one type of form vowel, except for ra and la cannot have a vocal e or o.

To end the the sound of syllables, can be done by adding a punctuation like arda Chandra, surang, bisah, cecek or adeg-adeg

3. Results and Analysis

Data image of the Balinese script on papyrus which is used in this research was taken from Lontar Wariga Gemet dataset obtained from the website http://www.archive.org that documented and published by the Documentation Center of Dinas Kebudayaan in Bali. This research used a sample image that consists of Balinese script papyrus image datasets which is distributed as a training image dataset of 70% and test image dataset of 30%.The testing process in this study were divided into two scenarios. The first scenario, we examine the success rate of LDA in classifying each Balinese script characters in a papyrus. second, we examine the success of romanization rules Balinese script. Figure 3 is the image of the printscreen system that used to reconstruct Balinese papyrus images.

Figure 3. Reconstruction Process

In the test stage of reconstruction with LDA, testing was conducted by counting the number of individual Balinese script character test that classified correctly, the number of Balinese script character test that is not recognized, the number of Balinese script character test that is recognized as other characters and the number of objects that are not characters in the test images that are recognized as the Balinese script. Values of the test results then calculated The percentages by dividing the sum with the total number of characters tested, are used the equations :

$$\text{Percentages} = \frac{\text{number of characters}}{\text{total number of characters}}\ x\ 100\% \quad (7)$$

Total number of character	15.120
Number of character test that is recognized	12.229
Number of character test that is not recognized	8.76%
Number of character test that is recognized but not character in Balinese script	903

Number of character test that is recognized as other characters	10.29%
Reconstruction success rate	80.88%

The Scenario test of Romanization Balinese script is done by counting the number characters are converted correctly according to the rules that have been prepared. In addition to calculating characters are converted correctly, also calculated the number of errors in the conversion and then do a comparison that number by the total number characters.

4. Conclusion

Based on this study, it can be concluded that the LDA method can classify Balinese script on papyrus with a good accuracy. LDA is able to distinguish each character of Balinese script even if just a little different. The percentage of success LDA method from the research that have been done in the amount of 80.88%. LDA classification process is highly dependent from the result of image segmentation Balinese script on papyrus. In this study, the process romanization of Balinese script able to perform translational very well. Romanization process already have the standard rules that making it easier in translation of Balinese script into Latin.
The next study that can be done is to spell checking results of romanization so that each word is generated based on the content Balinese script on papyrus have meaning. To complete the reconstruction of this papyrus, the result of a spell checking process can be translated back into Balinese script.

Acknowledgement

Appreciation is awarded to all the colleagues who have given a lot of help both in discussion of ideas until the trial implementation stage of this research. Not forgotten also the appreciation is delivered to the Department of Electrical and Computer Engineering, Udayana University which has supported in lending the laboratory equipments in this research.

References

[1] Tinggen, I Nengah. Pedoman Perubahan Ejaan Bahasa Bali. Singaraja: CV Rhika Dewata. 1994.
[2] Sudarma, Made. *Identifying Of the Space Color Cielab for the Balinese Papyrus Characters International Journal of Soft Computing*. 2015. 10 (3).
[3] Sudarma, Made, Agus Surya Darma. The Indentification of Balinese Scripts Character Based on Semantic Feature and K nearest Neighbor. *International Journal of Computer Applications (IJCA).* 2014: 91(1).
[4] Saad Bin Ahmed. Balinese Character Recognition Using Bidirectional LSTM Classifier. *Lecture Notes in Electrical Engineering (LNEE)*. 2016; 387: 201-211. DOI: 10.1007/978-3-319-32213-1_18.
[5] Lei Guo. Characters Feature Extraction based on Neat Oracle Bone Rubbings. *TELKOMNIKA*. 2013; 11(9): 5427-5434.
[6] M Boutaounte, Y Ouadid. Tifinagh Characters Recognition Using Simple Geometric Shapes. *Indonesian Journal of Electrical Engineering and Computer Science.* 2016; 3(1): 235–239.
[7] Kauleshwar Prasad, et al. Character Recognition Using Matlab's Neural Network Toolbox. *International Journal of u- and e- Service, Science and Technology*. 2013: 6(1).
[8] Sudarma, Made, Sutramiani. The Thinning Zhang-Shuen Application Method in the Image of Balinese Scripts on the Papyrus. *International Journal of Computer Applications (IJCA).* 2014; 91(1).
[9] Imam Habibi, Rinaldi. The Balinese Unicode Text Processing. *Indonesian Journal of Innovations in Soft Computing and Cybernetic Systems*. 2006: 1(1).
[10] http://www.unicode.org/charts/PDF/U1B00.pdf. *Unicode Balinese Script*. Cited: June 20, 2016.
[11] ShuangXu, Min Li, Yanqiu Cui. A Mixed Two-dimensional Linear Discriminate Method. *TELKOMNIKA*. 2013; 11(6): 3012-3019.
[12] Fisher RA. *The Use of Multiple Measurements in Taxonomic Problems*. Annals of Eugenics. 1936; 7(2): 179–188. DOI:10.1111/j.1469-1809.1936.tb02137.x.hdl:2440/15227.
[13] Hassan E *Searching OCR'ed Text: An LDA Based Approach.* Document Analysis and Recognition (ICDAR). International Conference on. 2011. DOI: 10.1109/ICDAR.2011.244.
[14] Kurt Z et al. *Linear Discriminant Analysis in Ottoman Alphabet Character Recognition*. Proceedings of the European Computing Conference. Lecture Notes in Electrical Engineering. 2009; 28: 601-607.
[15] Yuvika Dhillon. Character Recognition Using Neural Network and Linear Discriminant Analysis (LDA). *International Journal for Research and Science (IJFRS).* 2013; 2(3): 10-13.