The Identification of Balinese Scripts' Characters based on Semantic Feature and K Nearest Neighbor

Made Sudarma
Computer System and Informatics
Department of Electrical Engineering
Faculty of Engineering, Udayana University
Bukit Jimbaran Campus, Bali, Indonesia

I Wayan Agus Surya Darma
Magister Program of Elctrical Engineering,
Udayana University Graduate Program
Jl. PB Sudirman, Denpasar 80232
Bali - Indonesia

ABSTRACT
Papyrus script is a cultural heritage in Bali. As we know, that the papyrus is a cultural matter which is rich in valuable cultural values. Issues or problems encountered today is that the papyrus are not well maintained. Thus, many papyrus becomes damaged because it is not stored properly. Papyrus script was written using Balinese script's characters which having different features compared with Latin's characters. Balinese script can be recognized with feature extraction owned by each Balinese script. KNN is a classification algorithm based on nearest neighborhood. KNN can be used to classify Balinese script's features so that the test Balinese script's features which having nearest neighborhood value with the trained Balinese script's features will be recognized as the same Balinese script.

Keywords
Balinese scripts, Feature extraction, KNN

1. INTRODUCTION
The identification of hand-writing's characters presence in documents such as in letters, forms and other manuscripts has various applications in various fields [1]. Papyrus manuscript is an art created by ancient Balinese society. Characters presence in papyrus manuscript is Balinese script's characters which is ancient characters in Bali.

Balinese script can be recognized based on features owned by each character in Balinese script. Balinese script consists of the character *ha, na, ca, ra, ka, da, ta, sa, wa, la, ma, ga, ba, nga, pa, ja, ya, nya*, written in the form of script's character symbols. Feature which can be extracted from Balinese script is semantic feature [6]. Semantic feature is used to find out the semantic characteristic owned by Balinese writing characters such as the number of black strokes, loops, end nodes, characters' length, character's width, or open nodes of the characters.

The utilization of K Nearest Neighbor method in this research is used as a classification method to the feature owned by Balinese script. The result of classification process will be used as a reference in character recognition of Balinese script on papyrus manuscript.

2. FUNDAMENTAL THEORY
2.1 K Nearest Neighbor (KNN)
One of our ancestor's cultural legacies which have important value is ancient manuscripts. All over Indonesia it is known that there are lots of ancient manuscripts written in various scripts and languages. Most of the manuscripts still stored or owned by ordinary people. The others are existed in central and regional agencies, and traditional institutions. K Nearest Neighbor (KNN) is an interesting classification method in data mining which often called as lazy learning. K Nearest Neighbor algorithm is often used in classification and also used in the prediction and estimation. K Nearest Neighbor is included in instance-based learning that is training data is stored so that classification for new data can be classified by comparing the data most resembled with training set by nearest neighbor concept. A neighbor is considered nearest if having the shortest distance. K Nearest Neighbor stores all the training samples and postpones the learning until the new data should be classified, this is why it is called as lazy learning or slacker.

The implementation of KNN works based on the shortest distance from the test data to the sample data to determine its k nearest neighbor[4]. Sample data is projected into multiple dimension spaces, where each dimension represents feature of data. The space is divided into the sections based on sample data classification. A point in this space is marked with class c if class c represents the most found classification in k nearest neighbors from that point. To count the distance between new data and training set so that to obtain the distance of neighbor using distance function.

The steps to count K Nearest Neighbor method is:

1. To determine the K parameter (the amount of nearest neighbors).
2. To calculate the Euclidean distance quadrate (query instance) of each object towards sample data provided.
3. Subsequently to sort the objects into the group which having the shortest distance.
4. To collect the Y category (nearest neighbor classification).
5. By using the most majority of nearest neighbor category then query instance value that has been calculated can be predicted.

2.2 Balinese Scripts
The history of Balinese script is close related with the script development in India. Balinese script is derived from the language and script carried from India at the time of religion spreading of Hindu and Buddha era into Indonesia. The following is the sample of Balinese script writing.

Table 1. Hand-writing's characters of Balinese script

Ha		*La*	
Na		*Ma*	
Ca		*Ga*	
Ra		*Ba*	
Ka		*Nga*	
Da		*Pa*	
Ta		*Ja*	
Sa		*Ya*	
Wa		*Nya*	

2.3 Feature Extraction

Feature extraction is used to find out the patterns or features owned by each character which further process is doing classification for character recognition [5]. The outline is that a character has 3 characteristics:

1) The size of character's width and height. The size is taken from the average of each character inserted as a learning.

2) Parts of the character. It is a writing division into three areas which is upper part (ascender), middle part (main body) and lower part (descender). Afterward each part is taken of its feature using histogram to differentiate characters.

3) Stroke modeling. Stroke modeling uses a series of strokes (writing lines) to recognize character. A series of stroke is a group of points which is given number labels based on next neighbor's point direction stored inside the list which then its pattern being checked. The label given is as follows:

   a. Number 1 for an upward or downward direction.

   b. Number 2 for a direction to the left or right.

   c. Number 3 for a diagonal direction from upper right to the lower left.

   d. Number 4 for a diagonal direction from upper left to the lower right.

Table 2. Geometric Features.

Geometric features	Description	Geometric features	Description
	Open arches from 2700 to3600, consists of a series of label numbers 2, 3 and 1 respectively		Open arches from 2700 to 900, consists of a series of label numbers 1, 3, 4, 2, and 1 respectively
	Open arches from 1800 to 2700, consists of a series of label numbers 1,4 and 2 respectively		Close arches consists of a series of label numbers 4, 1, 3, 2, 4, 2, 3 and 2 clockwise direction respectively
	Open arches from 00 to 900, consists of a series of label numbers 2, 4 and 1 respectively		Slashes, consisting of a series of label numbers 3
	Open arches from 900 to 1800, consists of a series of label numbers 1, 3 and 2 respectively		Slashes, consisting of a series of label numbers 4
	Open arches from 00 to 1800, consists of a series of label numbers 2, 4, 1, 3 and 2 respectively		Vertical line, consisting of a series of label numbers 1
	Open arches from 900 to 2700, consists of a series of label numbers 1, 4, 2, 3 and 1 respectively		Horizontal line, consisting of a series of label numbers 2
	Open arches from 1800 to3600, consists of a series of label numbers 2, 3, 1, 4 and 2 respectively		

3. RESEARCH METHOD

3.1 Balinese Script's Features

Feature which can be extracted from Balinese script is semantic feature. Semantic feature which wanted to obtain from script's image is in the form of total stopping points, total rows and columns, total iterations or black dots forming a circuit, total horizontal lines and total vertical lines. The image of Balinese scripts resulting from segmentation is stored inside the matrix nxn prior to searching process of Balinese Script's features. The following is the stages in obtaining the features of Balinese Script's image based on semantic feature.

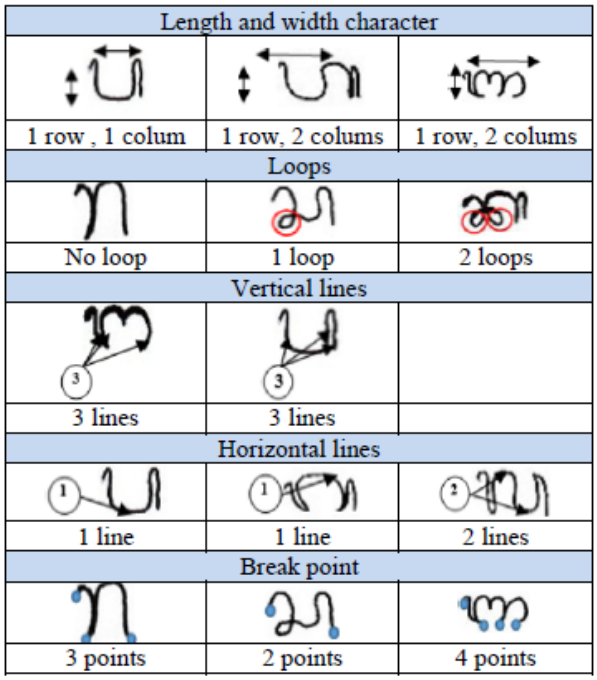

Figure 1. Semantic feature on Balinese script.

   a. Stopping Points
   Stopping points in Balinese script can be used as script's special features for recognition process of Balinese script. Each of Balinese script has various stopping points, the following is an example of stopping points in *Na* script.

Figure 2. Stopping points in *Na* script.

b. Length and Width of Character
The writing of Balinese script has various length and width. Length and width of the character is a semantic feature which can be used in character recognition. The following is an example of character's length and width of Balinese script Ha.

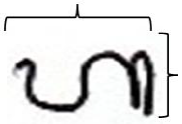

Figure 3. Length and Width of character *Ha*.

c. Number of Rows and Columns in Balinese script
The writing of Balinese script has the feature of total rows and columns of Balinese script's characters. The following is an example of rows and columns' length of script *Ha*.

Figure 4. Total Rows and Columns in Balinese script.

d. Loop
Total loop in Balinese script's character, is a feature which can be used in character recognition. The following is the writing of character *Ka* which has two loops.

Figure 5. Loop in character *Ka*.

e. Horizontal Lines
The writing of Balinese script's character has features in the form of total horizontal lines in Balinese script's character. The following is an example of Balinese script's writing *Ma* which has three horizontal lines.

Figure 6. Total of Horizontal lines in Balinese script *Ma*.

f. Vertical Lines
Vertical lines can be used as a feature which differentiates one Balinese script with another Balinese script. The following is an example of vertical line presence in the character of Balinese script *Ha*.

Figure 7. Total of Vertical lines in Balinese script *Ha*.

3.2 Classification of KNN

Implementation of KNN works based on the shortest distance from the test data to the sample data to determine its K nearest neighbor. Sample data is projected into multiple dimension spaces, where each dimension represents feature of data. This space is divided into the sections based on the classification of sample data. A point in the space is marked with class c if class c is a most found classification in k nearest neighbors from that point[2]. It calculates the distance between new data and training data so that it obtains the neighbor's distance using a distance function.

KNN classification is carried out on new data by using a distance function of Dynamic Time Warping (DTW). Distance function of DTW from two vectors *U* and *V* with the length of *m* and *n* is written with the formula as follows [2]:

$$DTW(U,V) = \gamma(m,n) \quad \text{.....(1)}$$

$$\gamma(m,n) = d_{base}(u_i, v_j) + \min\begin{cases}\gamma(i-1,j)\\ \gamma(i-1,j-1)\\ \gamma(i,j-1)\end{cases} \quad \text{.....(2)}$$

$$\gamma(0,0)=0, \gamma(0,\infty)=\infty, \gamma(\infty,0)=\infty$$
$$(i = 1,2,3..m;\ j = 1,2,3..n) \quad \text{.....(3)}$$

Table 3. The Result of Semantic Feature Extraction

Balinese script	Break points	Total length	Total width	Total loops	Total horizontal lines	Total vertical lines
Ha	3	2	1	0	2	3
Na	2	1	1	2	0	2
Ca	3	1	1	1	1	2
Ra	3	1	1	0	0	2
Ka	3	2	1	2	2	2
Ra	3	1	1	0	0	2
Da	2	1	1	1	0	2
Ta	3	2	1	2	2	2
Wa	3	1	1	0	2	2
Ma	5	1	1	1	0	1
Na	6	1	1	2	1	3
Ga	3	1	1	1	1	3
Ba	5	2	1	0	0	3
Da	2	1	1	2	0	2
Ya	4	2	1	0	1	1
Nya	5	2	1	0	1	0

For example there are two features namely *U = (1 8 5 1)* and *V = (2 1 8 6)*. The calculation of DTW for these two features is as follows:

Table 4a. The Calculation of DTW 1.

d	=d	cumulative	= cumulative	Remark
$(2-1)^2$	1	1 + 0	**1**	Minimum
$(2-8)^2$	36	36 + 1	37	
$(2-5)^2$	9	9 + 37	46	
$(2-1)^2$	1	1 + 37	47	

Table 4b. The Calculation of DTW 2.

d	=d	cumulative	= cumulative	Remark
$(1-1)^2$	0	0 + 1	1	Minimum
$(1-8)^2$	49	49 + 1	50	
$(1-5)^2$	16	16 + 37	53	
$(1-1)^2$	0	0 + 46	46	

Table 4c. The Calculation of DTW 3.

d	=d	cumulative	= cumulative	Remark
$(8-1)^2$	49	49 + 1	50	
$(8-8)^2$	0	0 + 1	1	minimum
$(8-5)^2$	9	9 + 1	10	
$(8-1)^2$	49	49 + 10	59	

Table 4d. The Calculation of DTW 4.

d	=d	cumulative	= cumulative	Remark
$(6-1)^2$	25	25 + 50	75	
$(6-8)^2$	4	4 + 1	5	
$(6-5)^2$	1	1 + 1	2	minimum
$(6-1)^2$	25	25 + 2	27	

Table 4e. The Result of DTW calculation based on the Calculation of U and V at table 4a up to 4d.

	2	1	8	6
1	1	1	50	75
8	37	50	1	5
5	46	53	10	2
1	47	46	59	27

4. RESULTS AND ANALYSIS

The experiments were performed using a script bali wianjana by extracting semantic features which are owned by the wianjana script. Based on the derived features, KNN classification is applied, the value of the features that have value to the neighborhoods closest training feature will be considered as the same script. Testing is done by extracting the semantic features which are owned by Balinese test, the following is the process of extraction of semantic features on Balinese.

Script Input:

1. The number of stopping points

Stopping point is obtained by detecting the black pixels using adjacency board membership.

P1	P2	P3
P8	X	P4
P7	P6	P5

Figure 8. Neighborhoods of pixels which includes stopping point.

Iteration process is done on the image Balinese character based board membership in order to obtain the number of stopping points on the image of Balinese. Here is a number derived from the stopping point Balinese Sa totaling two stopping points.

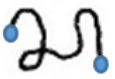

Figure 9. Stopping point on Sa Balinese character.

2. Total length and width of characters

Search length and width of the character alphabet letters bali bali by detecting pixels in the image. Once the location of the black pixel is found, the length and width of the character obtained by reducing the width of the final location of the initial width layout resulting in a number of lines of length and width of the character. Here is the length and width of a character that has a number of long *Sa* = 1 and width = 1.

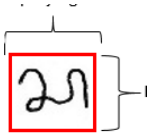

Figure 10. The length and width of Balinese character *Sa*

3. Number of loops

Search is then performed on the number of loops Balinese character *Sa*. The use of board membership as follows.

P1	P2	P3
P8	X	P4
P7	P6	P5

Figure 11. The initial search loop on the character *Sa*

In order to obtain the loop on *Sa* Balinese character image using the adjacency board membership as shown in figure 12.

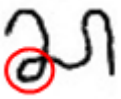

Figure 12. Loop script contained in *Sa*.

4. The number of horizontal lines

The horizontal line is obtained by using a board membership with adjacency as follows:

P1	P2	P3
P8	X	P4
P7	P6	P5

Figure 13. Ketetanggan pixels which includes a horizontal line.

Based on the adjacency of pixels found on Balinese horizontal line, as shown in figure 14.

Figure 14. The horizontal line on Balinese *Sa*

5. The number of vertical lines

The vertical line is obtained by using a board membership with adjacency as follows:

P1	P2	P3
P8	X	P4
P7	P6	P5

Figure 15. Neighborhoods of pixels which includes a vertical line.

Based on the neighborhoods of pixels found on Balinese vertical line, as in Figure 16.

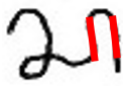

Figure 15. Vertical lines at Balinese *Sa*

Based on the search process semantic features that have made the Balinese script *Sa*, feature data is stored in the form of a matrix as shown in Table 5 below.

Table 5. Semantics feature Balinese script test

Balinese script	Break points	Total length	Total width	Total loops	Total horizontal lines	Total vertical lines
	2	1	1	1	1	2

Semantic features which have been obtained from the feature extraction process will be used as a special characteristic that distinguishes Balinese Balinese each other. So based on these semantic features can be identified by using a script KNN classification. Each column in the table declare semantic features contained in Balinese script.

5. CONCLUSION

The experiment result presents that 54 Balinese script images acquired by the success percentage 88.89%. It can be seen from the image of Balinese 54 tested, the introduction of changes can be made in 48 Balinese script images. The introduction of characters that do not fit due to the similarity of semantic features which are owned by the Balinese character, an example of the character and the character *Ca Sa*. But overall Balinese can be recognized well with the percentage of 88.89%.

Based on the results it can be concluded that the utilization of K Nearest Neighbor method is able to carry out the recognition of Balinese script based on semantic feature owned by each Balinese script. The development of the research which has been carried out in this research, can be maximized by adding the features used as reference in recognition, such as direction feature owned by Balinese script's characters. So that it can provide better recognition result on Balinese script's characters.

6. ACKNOWLEDGEMENTS

A great appreciation goes to colleague and everybody who has made valuable contributions in this study and their critical comments on this manuscript.

7. REFERENCES

[1] Aggarwal, Ashutosh, Rani, Rajneesh, Handwritten Devanagari Character Recognition Using Gradient Features, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5, ISSN: 2277 128X, 2012.

[2] Akbari. Mohammad, Eslami. Reyhaneh, Kashani. M. H., *Offline Persian Writer Identification Based on Wavelet Analysis*, International Conference on Bioinformatics and Biomedical Technology vol.29, 2012.

[3] Bhokse. Bhushan C., Thakare. Bhushan S., Devnagari Handwriting Recognition System using Dynamic Time Warping Algorithm, International Journal of Computer Applications, Volume 52– No.9, 0975 – 8887 August 2012.

[4] Elglaly. Yasmine, Quek. Francis Isolated Handwritten Arabic Characters Recognition using Multilayer Perceptrons and K Nearest Neighbor Classifiers, Computer Science Department, Virginia Polytechnic Institute and State University

[5] Putra. IKG Darma, Pengolahan Citra Digital, Andi No ISBN 9789792914436, 2010

[6] http://www.babadbali.com/aksarabali/books/ppebb.htm diakses terakhir tanggal 4 January 2014