Far East Journal of Mathematical Sciences (FJMS)

Published Online: August 2015
http://dx.doi.org/10.17654/FJMSOct2015_333_345
Volume 98, Number 3, 2015, Pages 333-345 **ISSN: 0972-0871**

THE INTELLIGENCE ALGORITHM FOR CHARACTER RECOGNITION ON PALM LEAF MANUSCRIPT

Narumol Chumuang[1], Preecha Khakham[1] and Makasak Ketcham[2]

[1]Department of Information Technology
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
e-mail: lecho20@hotmail.com
mitt_54784@hotmail.com

[2]Department of Information Technology Management
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
e-mail: mahasak.k@it.kmutnb.ac.th

Abstract

This paper proposes a new approach for Isan Dhamma handwriting on palm leaves character recognition system based on logistic model tree (LMT). The proposed system consists of three main processes. The first step is image preprocessing, feature extraction and Isan Dhamma handwriting recognition. The LMT consists of a standard decision tree structure with logistic regression functions at the leaves but high performance techniques for supervised learning and new method for classification in image processing. The amount of 300 images is used in the experiment where can be classified Isan Dhamma character effectively. The mean correct classified rate was 83.33%.

Received: April 8, 2015; Accepted: June 6, 2015
2010 Mathematics Subject Classification: 68T10, 68U10.
Keywords and phrases: Isan Dhamma, handwriting, recognition, LMT (logistic model tree).
Communicated by K. K. Azad

1. Introduction

Isan Dhamma are the ancient characters in northeastern of Thailand that look round of the Lanna, see in Figure 1. During the 22nd-24th Buddhist century. There has been evidenced that Isan Dhamma manuscript were used to written record important stories about history and Buddhist scriptures such as the bible, literature, witchery, herbal medicines and treatments, and others on the surface of stone or on palm leaf shown in Figure 2. The palm leaves were recorded, these were called "*Bai-Lan book*" [1].

Figure 1. Example of Isan Dhamma characters.

They had many important knowledge. Currently there are a small amount of readers who can read and understand Isan Dhamma [2]. Nowadays, it seems to have decreased which make knowledge lost. These were an anciently heritage that have sanctity in the historic. The oldest Isan Dhamma manuscript was found at Sri Khunmuang temple in Nongkai province in Thailand [3].

Figure 2. Palm leaves manuscripts.

The recognition techniques can apply to extract the knowledge out of palm leaves as a digital document for keeping. It can keep the knowledge and wisdom of the ancestors that accumulate so long. It also led them to use the knowledge for the benefit of future offspring.

Preprocessing and feature extraction are factors that increase the efficiency in recognition [4]. The research in many years found that the various commonly used as projection histogram [5, 6], template matching [7-9], zoning [10, 11], gradient feature [12], grid feature and global feature [13-16]. Since all the documents used in this research are very old. The fibers of the palm leaf, water stains and mold were lot of noise then this paper will not presented to obviate noise but we offer outstanding feature extraction and the high performance algorithm. The result shown LMT algorithm was a good performance in character recognition.

2. Related Researches

2.1. Isan Dhamma character

During the past with attention and presented a variety of methods. In order to help the reader understand Isan Dhamma was easier for the common man. The aim was to acquire knowledge that was recorded on palm leaves to continue. The related research works are as follows.

Yuttaponraneepinit and Wantanapa [17] compared with four techniques for segmentation that were global threshold, edge detection, contour following and local threshold. The experimental results shown that local threshold technique overcome the other three techniques in all cases.

Thammano and Trikeawcharoen [18] presented method for classification old Asian script handwritten from 23 educated Buddhist monks without feature extraction process. With the use of Hausdorff distance as the similarity measure and the new mutation operator, the proposed evolutionary algorithm can achieve. The average performance was 64.85%.

Promsoda and Seresangtakul [19] proposed the translation of Isan Dhamma alphabets to the Thai language by using a context free grammar. A dictionary was created by tries data structure. The hallmark of structure is

easy to control a number of different lengths and convenient way to store and access data. The longest matching technique used for translation. The experimental results illustrated the translation accuracy of 70%.

Wijitchareon and Seresangtakul [20] presented Isan Dhamma handwritten character classification and recognition by using *k*-fold cross validation for zoning based histogram projection both of horizontal and vertical with back propagation neural network. The experimental results showed the proposed method performed higher accuracy.

The experiment in this paper used five Isan Dhamma characters. The comparisons of Thai characters with Isan Dhamma are shown in Table 1.

Table 1. Compare Thai characters with Isan Dhamma

No.	Thai character	Isan Dhamma character
1	ก	
2	ด	
3	ต	
4	ป	
5	พ	

2.2. Logistic model tree

LMT is high performance techniques for supervised learning tasks, both for the prediction of nominal classes and numeric values [21]. The elementary consists of a standard decision tree structure with logistic regression functions at the leaves. As in the common decision trees, a test on one of the attributes is associated with every inner node. This algorithm is the combination between of linear logistic regression and tree induction. In the past produces low variance high bias and the later produces high variance low bias. These two techniques were combined into learner which depends upon simple regression models if only little and/or noisy data is present. It adds more complex tree structures if enough data is available to such

structures. Thus, logistic model trees are the decision trees having linear regression model at leaves.

A logistic model tree consists of a tree structure that is made up of a set of inner or non-terminal nodes *N* and a set of leaves or terminal nodes *T*. Let $S = D1 \times \cdots \times Dm$ denote the whole instance space, spanned by all attributes $V = v1, ..., vm$ that are present in the data. Then the tree structure gives a disjoint subdivision of S into regions S_t and every region is represented by a leaf in the tree are equation (1) [21].

$$S = \bigcup_{t \in T} S_t,\ S_t \cap S_{t'} = \theta \text{ for } t \neq t'. \tag{1}$$

The leaves $t \in T$ have an associated logistic regression function f_t instead of just a class label. The regression function f_t takes into account an arbitrary subset $V_t \subset V$ of all attributes present in the data, and models the class membership probabilities as equation (3).

$$P_r(G = j \mid X = x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}}, \tag{2}$$

where

$$F_j(x) = \alpha_0^j + \sum_{k=1}^{m} \alpha_{\upsilon_k}^j \bullet \upsilon, \tag{3}$$

or, equivalently,

$$F_j(x) = \alpha_0^j + \sum_{k=1}^{m} \alpha_{\upsilon_k}^j \bullet \upsilon_k \tag{4}$$

if $\alpha_{\upsilon_k}^j = 0$ for $\upsilon_k \notin V_t$. The model represented by the whole logistic model tree is then given by

$$f(x) = \sum_{t \in T} f_t(x) \bullet I(x \in S_t), \tag{5}$$

where $I(x \in S_t)$ is 1 if $x \in S_t$ and 0 otherwise.

In this paper, Isan Dhamma character recognition on Palm leaves using LMT is proposed. LMT's perform competitively with other state-of-the-art classifiers such as boosted decision trees while being easier to interpret. Although, LMTs want to use time for building the tree structure which mostly to the cost of building the logistic regression models at the nodes. But this algorithm has uncomplicated structure then it is easily understood. This technique shown experiment results are efficiency accurate for Isan Dhamma.

3. Recognition System

The systems are introduced into three main processes. The first is preprocessing, feature extraction and recognition. In the experiment, the dataset are divided into two groups for training and testing. It also provides the results are reliable then shown the graph of the cross validation as well. The block diagram of the system is given in Figure 3.

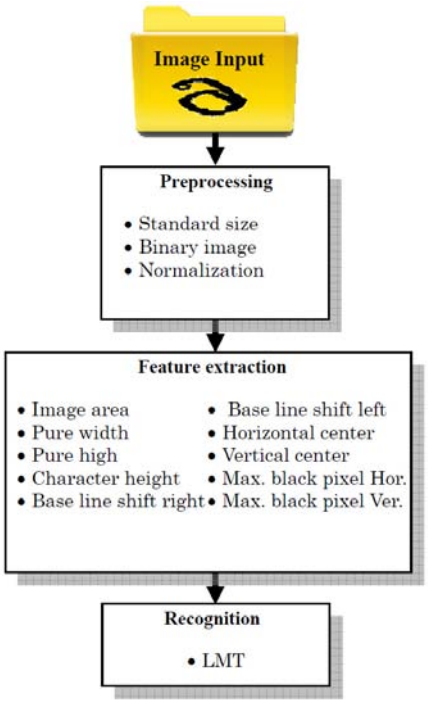

Figure 3. Block diagram of the system.

3.1. Preprocessing

This section is the first step of the system. Both of dataset for training set and testing have to do following.

Standard size. These images in database have different size. For solve this problem then the system determined standard size 80×160 pixels for every image to input, see Figure 4.

Binary image. The system converted all images to black and white images. Some input images have different line color. The only one reason has to convert because we emphasize the pattern of line for speed in computing.

Normalization. Because of handwriting is a person's behavior. The weight of the wrist and fingers are the size of the line in a different then the characters are different too. The devices used in writing, it may be different in the size of the line then we need skeleton.

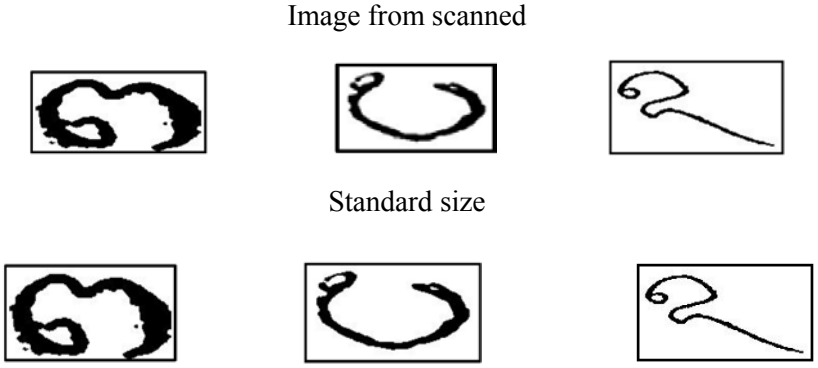

Figure 4. Example standard size on image.

Figure 5. Skeleton image.

3.2. Feature extraction

The system used global feature that provided detail to extract outstanding feature from Isan Dhamma character. All of extracted features are in this section came the inputs of training dataset.

Image area. The image area are the counting the number all of black pixels which belong of the image. This feature provides detail of the density of the image [21].

Image ratio. The image height-to-width ratio is obtained by dividing image height to image width.

Pure width. This feature means the net width of the image size excluding the background.

Pure height. The height of the image by the net without background [23].

The image amenable. The image amenable to the vertical center of left and right to provide the direction of image.

Maximum horizontal histogram and maximum vertical histogram. The horizontal histogram is calculated for each row and the row which has the highest value is taken as maximum horizontal histogram. The vertical histogram is calculated for each column and the column which has the highest value is taken as maximum vertical histogram.

Horizontal and vertical center of the image. These are calculated using the formulas in equations (6), (7) [13].

$$C_x = \frac{\sum_{x=1}^{x_{\max}} x \sum_{y=1}^{y_{\max}} b[x,\ y]}{\sum_{x=1}^{x_{\max}} \sum_{y=1}^{y_{\max}} b[x,\ y]}, \tag{6}$$

$$C_y = \frac{\sum_{y=1}^{y_{\max}} y \sum_{x=1}^{x_{\max}} b[x,\ y]}{\sum_{x=1}^{x_{\max}} \sum_{y=1}^{y_{\max}} b[x,\ y]}. \tag{7}$$

4. Character Database

Ours experiment, we used Isan Dhamma character on palm leaves to input image by scanning into database. As the palm leaves are very old. Scanning is to be done with carefully so that the input process requires much time.

In this paper, training and testing of recognition system 300 images from five characters of Isan Dhamma were used in experiment. For the training system 180 images were used and 120 images were used testing system.

5. Experiment and Result

The recognition section consists of two parts, training and testing, respectively, which is accomplished by LMT.

Training part. Isan Dhamma recognition is a multi-class problem. Since the LMT classifier for building logistic model trees which are classified trees with logistic regression functions at the leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values. One of the strengths of LMT is that it produces small, accurate trees (it uses the CART pruning mechanism). The parameters training of LMT in the system are displayed in Table 2. The tree structure was shown in Figure 6.

Table 2. Parameter of LMT training

Parameter	Value
Number of folds	100
Minimum of instance at which a node for split	15
Iterations for LogitBoost	15
Training time	No limit

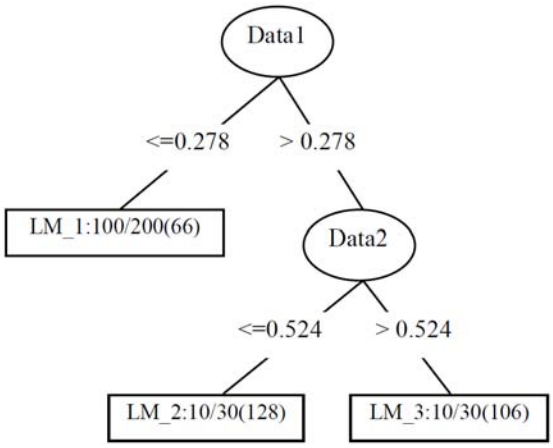

Figure 6. Isan Dhamma classification tree constructed by the LMT algorithm.

The validation performance Isan Dhamma system by using LMT algorithm shown the line graph with *k*-fold cross validation, see in Figure 7.

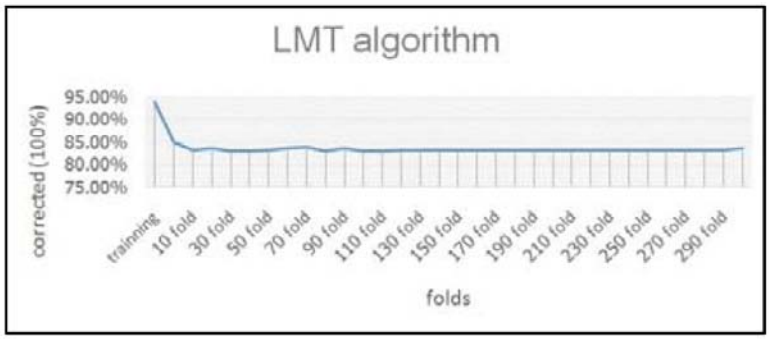

Figure 7. The cross validation performance of LMT.

Testing part. The system has been tested for its accuracy and effectiveness on 120 images. The experimental results shown that the proposed system can be recognized. The experimental results shown that the LMT algorithm can be classified five Isan Dhamma characters have the dataset 300 images used in this experiment was rated as very effective. The accuracy of the results values are shown average 83.33%. In experimental was tested 100 times by dividing the training set 60% and the test set 40%.

Table 3. Result in classification

No.	Character	Result	Accuracy rate(%)
1	ก		83.47
2	ด		77.61
3	ต		85.71
4	ป		85.19
5	พ		84.67

6. Conclusion

This paper proposed a new method to recognition Isan Dhamma character on palm leaves by using logistic model tree. Two excellent methods in this paper for made the accuracy performance was satisfactory that are good feature extraction and intelligence algorithm. The strengths of the LMT algorithm are the decision tree structure is smaller which allow for speed up processing time but high performances. Next this system used global feature from character image for extracted the overall appearance of the image to recognition.

References

[1] T. Punnotok, Ancient Eastern Characters, Humanities the Dhamma Characters and Thai Noy, Siampress Management, Bangkok, 1990, pp. 54-60.

[2] S. Wannaudon and T. Songnok, Attitudes towards learning the Isan Dhamma alphabet of students in Mahachulalongkorn Rajavidyalay University Mahasarakham Academic Service Unit, Journal of Humanities and Social Sciences University 31(2) (2012), 109-120.

[3] K. Khonwai, Isan Dhamma characters, Thai Culture Magazine 39(4) (2002), 60-64.

[4] C. Karnchanapusakij, Ph. Suwannakat and W. Rakprasertsuk, Online handwriting Thai character recognition, IEEE International Conference on Computer Graphics: Imaging and Visualization, 2009, pp. 323-328.

[5] N. Otsu, A threshold selection method form gray-level histogram, IEEE Transactions on System, Man and Cybernetics 9(1), (1979), 62-66.

[6] Konstantinos Rapantzikos, Yannis Avrithis and Stefanos Kollias, Dense saliency-based spatiotemporal feature points for action recognition, Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June, 2009, pp. 1-8.

[7] Eli Shechtman, Michal Irani, Matching local self-similarities across images and videos, Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, June, 2007, pp. 1-8.

[8] Ju Sun, Xiao Wu, Shuicheng Yan, Loong-Fah Cheong, Tat-Seng Chua and Jintao Li, Hierarchical spatio-temporal context modeling for action recognition, Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June, 2009, pp. 1-8.

[9] Christian Thurau and Vaclav Hlavač, Pose primitive based human action recognition in videos or still images, Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June, 2008, pp. 1-6.

[10] M. Blumenstein, B. K. Verma and H. Basli, A novel feature extraction technique for the recognition of segmented handwritten characters, 7th International Conference on Document Analysis and Recognition (ICDAR'03), Edinburgh, Scotland, 2003, 137-141.

[11] O. V. Ramana Murthy and M. Hanmandlu, Zoning based Devanagari character recognition, Int. J. Comput. Appl. 27(4) (2011), 21-25.

[12] G. Dimauro, S. Impedovo, R. Modugno, and G. Pirlo, Numeral recognition by weighting local decisions, Proc. Int. Conferences Document Analysis and Recognition, Edinburgh, U.K., Aug., 2003, pp. 1070-1074.

[13] A. Negi, K. N. Shanker and C. K. Chereddi, Localization, extraction and recognition of text in Telugu document images, Proc. Int. Conferences Document Analysis and Recognition, Edinburgh, U.K., Aug., 2003, 1193-1197.

[14] H. Baltzakis and N. Papamarkos, A new signature verification technique based on a two-stage neural network classifier, Pergamon, 2001, 95-103.

[15] N. Chumuang and M. Ketcham, Handwriting Thai signature recognition system based on multilayer perceptron and radial basis network, International Conference on Advanced Computational Technologies and Creative Media (ICACTCM'2014), 2014, pp. 44-49.

[16] Naruemol Chumuang and Mahasak Ketcham, Intelligent handwriting Thai Signature Recognition System based on Artificial Neuron Network, TENCON 2014-2014 IEEE Region 10 Conference, IEEE, 2014, pp. 1-6.

[17] V. Yuttaponraneepinit and S. Wantanapa, Palm Leaf Manuscript's Isan Dhamma Character Segmentation with Digital Image Processing, Diploma in Computer Science and Technology Faculty, Thammasart University, 2000.

[18] A. Thammano and S. Trikeawcharoen, Old southeast asian script recognition using evolutionary algorithm, IEEE Computer Society 3rd International Conference on Natural Computation (ICNC'2007), 2007, pp. 566-570.

[19] S. Promsoda and P. Seresangtakul, Isan Dhamma alphabet to Thai languages translation, The 7th National Conference on Computing and Information Technology (NCCIT'2011), 2011, pp. 503-508.

[20] A. Wijitchareon and P. Seresangtakul, A hybrid features extraction for Isan Dhamma handwritten character recognition using neural network, 4th Conference on Knowledge and Smart Technology (KST), 2012, pp. 45-52.

[21] N. Landwehr, M. Hall and E. Frank, Logistic Model Tree, ECML, Springer-Verlag Berlin Heidelberg, 2003, pp. 241-252.

[22] Chayut Wiwatcharakoses and Karn Patanukhom, Two-stage recognition for printed Thai and English characters using nearest neighbor and support vector machine, IEEE International Conference on Signal Image Technology and Internet-based Systems, 2013, pp. 71-78.

[23] J. P. Drouhard, R. Sabourin and M. Godbout, A neural network approach to offline signature verification using directional PDF, Pattern Recognition, September, 1996, 415-424.