Lanna Dharma Printed Character Recognition using *k*-Nearest Neighbor and Conditional Random Fields

Chutima Chueaphun[1], Atcharin Klomsae[1], Sanparith Marukatat[2] and Jeerayut Chaijaruwanich[1]
[1]*Department of Computer Science, Faculty of Science, Chiang Mai University, Chiang Mai, 50200, Thailand*
[2]*National Electronics and Computer Technology Center, Pathumthani, 12120, Thailand*

Keywords: Lanna Dharma, Character Recognition, *k*-Nearest Neighbor, Conditional Random Fields.

Abstract: For centuries, in the North of Thailand, many books of Lanna Dharma characters had been printed. These books are the important sources of the knowledge of ancient Lanna wisdom. At present, the books are found old and damaged. Most of characters are rough and not clear according to its early printing technology at that time. Moreover, some sets of characters are relatively very similar which cause the difficulty to recognize them. This paper proposes a Lanna Dharma printed character recognition technique using *k*-Nearest Neighbor and Conditional Random Fields. The accuracy of recognition rate is about 82.61 percent.

1 INTRODUCTION

Currently, there are many optical character recognition (OCR) researches which allow conversion of the text in scanned images into the machine-encoded text. OCR systems are available in many languages such as English, Japanese, Chinese, Arabic, Thai, etc. However, there are not yet any for Lanna Dharma character.

Hundreds of years ago, Lanna language was widely used in the northern part of Thailand during the time of the Lanna kingdom, which was founded in 1259. The Lanna Dharma character is a descendant of the old Mon character like Lao and Burmese characters. Since 1892, the typed Lanna Dharma character was first printed as books including history, medicine, literature and Buddhism. The Lanna Dharma books are the important sources of Lanna regional knowledge. However, after the invading of Ayutthaya kingdom from the central of Thailand, the Central Thai language became the official language learned in school. Now a day, the Lanna Dharma character has almost been forgotten. There are now only few people, usually old ones, who can read it. In addition, most of the Lanna Dharma books are lost and destroyed. Therefore, Lanna Dharma printed character recognition will help to preserve the ancient Lanna knowledge. Furthermore, the knowledge can be delivered to general public with electronically retrievable.

The Lanna Dharma writing system has no white-space between each word, but there is a white-space at the end of a clause or sentence. Lanna Dharma word consists of consonants, vowels, and tones at different levels. Specifically, many consonants have their alternate form when they directly follow other consonant. We distinguish the types of characters for this paper into eight different types as shown in Table 1.

The writing example of Lanna Dharma word is shown in Figure 1. It can be written by organizing the consonants or the middle vowels in level 1; the upper vowels are in level 2; the tones are in level 2 or 3; the final consonants are in level 4 and the lower vowels are in level 4 or 5.

In this paper, we mainly focus on the solution to the problem of distinguish character belonging to confusion sets. Indeed, Lanna Dharma printed character have relatively similar patterns which cause the recognition error. Figure 2 shows the example of confusion sets. We propose the use of *k*-Nearest Neighbor (*k*-NN) to firstly classify the class of character images. Then, sequence of character classes which is the output of *k*-NN is reclassified again by Conditional Random Fields (CRFs). CRFs are the conditional undirected graphical models, which model the conditional probabilities of character sequence. Therefore, it is expected to resolve the problem of *k*-NN for confusion sets and improve the final character recognition.

Chueaphun C., Klomsae A., Marukatat S. and Chaijaruwanich J..
Lanna Dharma Printed Character Recognition using k-Nearest Neighbor and Conditional Random Fields.
DOI: 10.5220/0004112801690174
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 169-174
ISBN: 978-989-8565-29-7

Table 1: Types of Lanna Dharma characters.

Tag	Type	Example
SC	Special Consonant	
CO	Consonant	
TC	Transformation Consonant	
UV	Upper Vowel	
LV	Lower Vowel	
MV	Middle Vowel	
TO	Tone	
NL	Number	

Figure 1: Writing example of Lanna Dharma word.

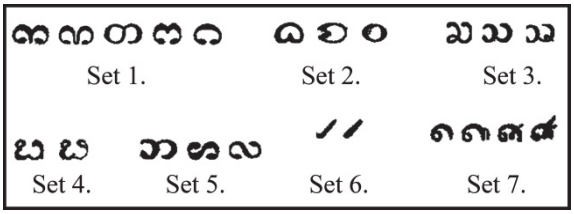

Figure 2: Example of confusion sets.

2 RELATED WORKS

Methods and techniques for OCR are various. Selecting which method and technique depends on the characteristics of characters and documents. Some methods are appropriate with some typed of characters while some are not appropriate with others.

The *k*-NN algorithm is a very well known approach. It is the simplest approach of all machine learning algorithms. A lot of works in character recognition use the *k*-NN classifier. In (Holambe et al., 2010), they provided a comparative study of Devanagari handwritten and printed character and numerals recognition using Nearest-Neighbor classifiers. This method used gradient and curvature based feature extraction method and compared all the nearest neighbor classification methods. In (Alkhateeb et al., 2009), *k*-NN was used for off-line handwritten Arabic word recognition. The word was segmented and divided into overlapping blocks. Absolute mean values were computed for each block of segmented words which constituted a feature vector. Finally, the resulting feature vectors were used to classify the word.

CRFs were introduced by Lafferty et al. (2001). A conditional random field is simply a conditional distribution $p(y|x)$ with an associated undirected graphical model structure. Within this model dependencies among the input variables x do not need to be explicitly represented, affording the use of rich, global features of the input. For example, in natural language tasks, useful features include neighboring words and word bigrams, prefixes and suffixes, capitalization, membership in domain-specific lexicons, and semantic information (Sutton and Mccallum, 2007). Moreover, CRFs have been applied to many domains, including computer vision, bioinformatics and word recognition. In word recognition, the first research, which CRFs were applied for handwriting recognition, was proposed in (Feng et al., 2006). CRFs were used to recognize the entire word without character segmentation and they formulated recognition problem as a problem of labeling observation sequences on a large-vocabulary of words. Similarly, the word recognition using CRFs was presented in (Shetty et al., 2007), the method was based on segmentation of word image into characters and identification of lexicon with the highest probability. Moreover, CRFs were applied to the same problem in other researches such as in (Zhou et al., 2009).

3 THE PROPOSED METHOD

In this section, the Lanna Dharma printed character recognition system is presented. It is comprising of four main processes including: preprocessing, segmentation, classification with *k*-NN, and reclassification with CRFs. The whole process is illustrated in Figure 3.

3.1 Preprocessing

In our experiments, we have collected document images from E-60 Lanna Printed E-book Project (Chaijaruwanich, 2010) which is a collection of 60 ancient Lanna books, printed between 1907-1947, including the legends, poems, laws, etc. The document images were scanned at 300 dpi and

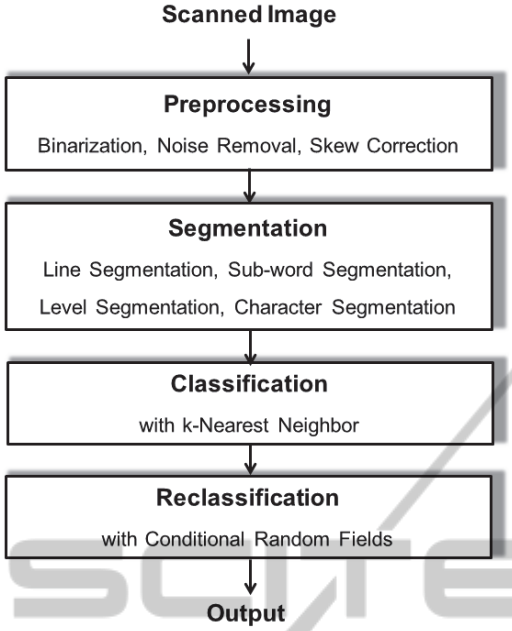

Figure 3: Lanna Dharma printed character recognition process.

stored as gray-level images. After document scanning, preprocessing operations will be applied to the document images to enhance the images to be the suitable format. Preprocessing is divided into three steps:

- Binarization
- Noise removal
- Skew correction

In the binarization step, Otsu's method (Otsu, 1979) is used for converting grey level image to a binary. In the noise removal step, noise is removed by connected-component labeling which detects the dusty pixels in the document images. If the size of pixels are smaller than the defined threshold, they are defined as noise and would be removed. Finally, in the skew correction step, we detect the skew angle of the document image using Hough Transform (Duda and Hart, 1972). Then, skew correction is done by rotating the document image referring to the skewed angle.

3.2 Segmentation

After finishing preprocessing, we apply the segmentation process which is divided into four steps:

- Line segmentation
- Sub-word segmentation
- Level segmentation
- Character segmentation

First, we employ horizontal projection profile to segment a page into separated lines. Second, vertical projection profile is employed to segment the lines by empty spaces into sub-words. Then, we use horizontal projection profile again to find upper and lower positions of a middle level for the level segmentation. In this paper, we consider three levels: upper level (i.e. levels 2, 3 in Figure 1), middle level (i.e. level 1 in Figure 1), and lower level (i.e. levels 4, 5 in Figure 1). Finally, sub-words are segmented into sequence of characters by using connected component labeling and center of mass techniques. Figure 4 shows an example of segmentation. As a result, the output from this process is a sequence of character images as shown in Figure 5 which is ordered by writing system of Lanna Dharma word.

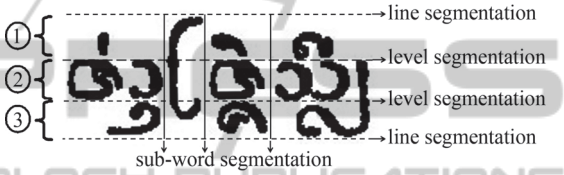

Figure 4: Example of segmentation, ① refers to upper level, ② refers to middle level, and ③ refers to lower level.

Figure 5: Sequence of character images ordered by writing system of Lanna Dharma word.

3.3 Classification with *k*-Nearest Neighbor

Generally, a classifier in built upon a sample of character images called the training set. The classifier decision is based on feature vector extracted from the character image.

3.3.1 Feature Extraction for *k*-NN

In this paper, height, width, aspect ratio, area and level from original size of character images are used as a feature vector. In addition, the character images are normalized into 36*36 pixels and they are divided into 9*9 cells, the ratios of number of the black pixels and area of each cell are also used as features. So in total we have 86 features for each character image.

3.3.2 Classification with *k*-NN

For *k*-NN classification, training patterns are data

points in *d*-dimensional space, where *d* is the number of features. An unlabelled test pattern is another point within the same space and classified by considering the most frequently occurring class among its *k*-most similar training patterns (Duda and Hart, 1973). For our experiments, we defined *k* is 3 and the similarity measure for *k*-NN classification is the Euclidian distance metric, defined between feature vectors *x* and *y* as:

$$dis(x,y) = \sqrt{\sum_{i=1}^{f}(x_i - y_i)^2} \quad (1)$$

where *f* represents the number of features, smaller distance values represent greater similarity.

3.4 Reclassification with Conditional Random Fields

We formulate a recognition problem as a problem of labeling observation sequences as in (Feng et al., 2006), and (Shetty et al., 2007). Generally, *x* is a random variable over data sequences to be labeled, and *y* is a random variable over corresponding label sequences. All components y_i of *y* are assumed to range over a finite label alphabet Y. In this paper, *x* range over Lanna Dharma character sequences and *y* range over labels of those sequences, with Y is the set of possible Lanna Dharma character symbols as in Table 1.

3.4.1 Feature Selection for CRFs

We use the following features: type of character as defined in Table 1, level of character as defined in section 3.2, and word boundary which are B (Beginning character of word) and I (Internal character of word). The example of Lanna Dharma character sequence, their features, and their true label are shown in Figure 6.

3.4.2 Classification with CRFs

The random variables *x* and *y* are jointly distributed, but in a discriminative framework as CRFs, we construct a conditional model $p(Y=y|X=x)$ from paired observation and label sequences. The probability of a label sequence $y = y_1, y_2,...,y_t$ given an observation sequence $x = x_1,x_2,...,x_t$ can be written as

$$p_\theta = p(y|x) = \frac{1}{Z(x)}\exp\left(\sum_{t=1}^{T}F(y,x,t)\right) \quad (2)$$

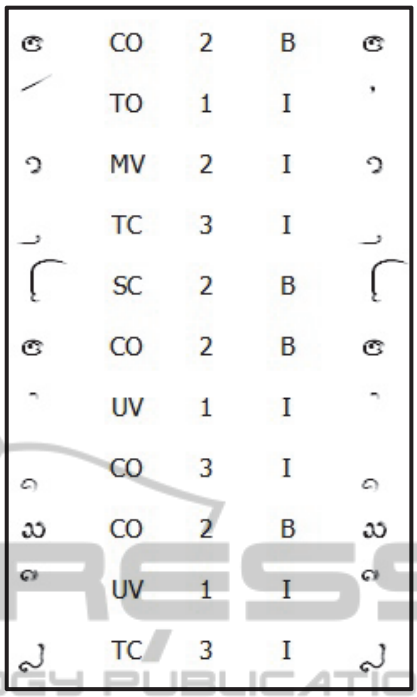

Figure 6: Example of Lanna Dharma character sequence, their features and their true label. The first column is the output from k-NN classification, the second column is the type of character, the third column is the level of character, the fourth column is the word boundary, and the fifth column is the true Lanna Dharma character label.

where

$$Z(x) = \sum_{y'} exp\left(\sum_{t=1}^{T}F(y',x,t)\right) \quad (3)$$

θ is the model parameter, *Z(x)* is a normalized factor which sums all cases of label y' and *F(y,x,t)* is the sum of feature frequency at position *t*.

We define the feature function of the entire Lanna Dharma character sequences and their features *x* and the corresponding Lanna Dharma character label *y* at position *t* as

$$F(y,x,t) = \sum_{l',l,j}\lambda^j_{l',l}f^j_{l',l}(y_{t-1},y_t,x) + \sum_{l,j}\beta^j_l\, g^j_l(y_t,x) \quad (4)$$

where $f^j_{l',l}(y_{t-1},y_t,x)$ is the transition feature function of Lanna Dharma character sequences and their features *x* and label *y* at position *t*-1 and *t* equal to l' and *l*, orderly in the whole states of Lanna Dharma character labels. $g^j_l(y_t,x)$ is the state feature function of the Lanna Dharma character label at position *t* of the Lanna Dharma character

sequence y and their features x. $\lambda^j_{l',l}$, and β^j_l are feature weights associated respectively with $f^j_{l',l}$ and g^j_l estimated from the training data. Let $\theta = (\lambda, \beta)$.

We define

$$z_j(x,t) = \begin{cases} 1 \text{ if } x \text{ at position } t-k \text{ to } t+k \\ \quad \text{is equal to } (n,s) \\ 0 \text{ otherwise} \end{cases} \quad (5)$$

$$f^j_{l',l}(y_{t-1}, y_t, x) = \begin{cases} z_j(x,t) \text{ if } y_{t-1} = l' \text{ and } y_t = l \\ 0 \quad \text{otherwise} \end{cases} \quad (6)$$

$$g^j_l(y_t, x) = \begin{cases} z_j(x,t) \text{ if } y_t = l \\ 0 \quad \text{otherwise} \end{cases} \quad (7)$$

From equation (5), $z_j(x,t)$ represents whether x at position t - k to t + k consisting of the Lanna Dharma character subsequence n and the Lanna Dharma character feature s. We consider a reasonable local information pattern of k = ± 2 Lanna Dharma character surround the considering position t. In equation (6), and (7), l and l' are Lanna Dharma character labels.

CRFs prediction is performed by finding the most probable label sequence y^* from the training model given the observation or Lanna Dharma character sequences and the Lanna Dharma character features x from the testing data.

$$y^* = argmax_y p_\theta(y|x) = argmax_y \exp\left(\sum_{t=1}^{T} F(y,x,t)\right) \quad (8)$$

The Viterbi algorithm (Lafferty et al., 2001) is used as a dynamic programming algorithm to generate the inferences in CRFs. This is to find the most probable label y^* for the Lanna Dharma character sequence and their features x with highest probability value.

The graphical model of our proposed CRFs are shown in Figure 7. We consider a range of two characters forward and backward from the indicated position t. These relationships are also used for training and testing data formats.

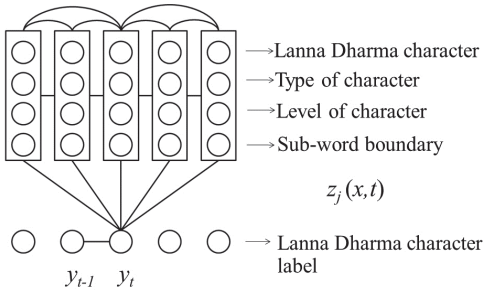

Figure 7: Graphical model of our CRFs model.

In our experiment, CRFs are implemented using CRF++ 0.53 (Kudo, 2005) which is an open source program providing a generalized CRFs learning and testing platform. CRF++ have been successfully used in a lot of reseach such as (Shoombuatong et al., 2011) and (Subpaiboonkit et al., 2012).

4 EXPERIMENTS

Generally, the recognition system needs a large database to train and test the system. Therefore, we select Lanna poem documents which are parts of the E-60 Lanna Printed E-book Project (Chaijaruwanich, 2010) to identify a scope of Lanna word, and collect a small database to train and test the Lanna Dharma printed character recognition system. The documents are selected with each page consists of 900-1,200 characters and they vary in levels of noise.

30 pages of document are randomly sampled for training k-NN classifier. The training dataset consists of 145 classes, we randomly select 30 sample images for each class. Totally, we have 4,350 sample images of training data for classification with k-NN. Samely, training data for CRFs classifier, 30 pages of document are randomly sampled, including 1,983 sequences or 33,721 characters.

To test our method, we randomly select 10 pages of documents, including 714 sequences or 11,425 characters. We compare our method with recognition using k-NN without CRFs. Table 2 shows the accuracy result of recognition using k-NN without CRFs, and using both k-NN and CRFs.

Table 2: The recognition accuracy rate.

Classifier	Recognition Accuracy Rate (%)
k-NN	78.27
k-NN and CRFs	82.61

Recognition accuracy of our method in Table 2 is higher than recognition using k-NN without CRFs. The recognition errors from our method may be caused by two sources: the broken and touching characters, and too small training dataset for CRFs. Broken and touching characters error refers to the case that the broken and touching characters are incorrectly segmented, then they are the causes of recognition error. Small training dataset for CRFs may cause a prediction error for unseen patterns.

5 CONCLUSIONS

In this paper, we propose Lanna Dharma printed character recognition and focus on solving the problem of confusion sets. After the preprocessing and segmentation processes, we classify Lanna Dharma printed character using *k*-Nearest Neighbor. The output from classification using *k*-Nearest Neighbor is reclassified again by CRFs. To use CRFs, we define Lanna Dharma character sequences which are output from classification using *k*-Nearest Neighbor as the sequence labeling task, and we construct a suitable template for the CRFs to train and test the data. The experimental results show that our method increases the recognition accuracy.

ACKNOWLEDGEMENTS

The research described in this paper was supported by the Graduate School Chiang Mai University, Chiang Mai, Thailand.

REFERENCES

Alkhateeb, J. H., Khelifi, F., Jiang, J. and Ipson, S. S. (2009). A new approach for off-line handwritten Arabic word recognition using KNN classifier. *Proc. ICSIPA*, 191 -194.

Chaijaruwanich, J. (2010). E-60 Lanna Printed Book Project. Retrieved from: http://www.lannalit.org

Duda, R. O. and Hart, P. (1972). Use of the hough transformation to detect lines and curves in pictures. *Comm. ACM*, 11–15.

Duda, R. O. and Hart, P. (1973). *Pattern classification and scene analysis*. Wiley, New York.

Feng, S., Manmatha, R. and Mccallum, A. (2006). Exploring the use of conditional random field models and hmms for historical handwritten document recognition. *Proc. 2nd DIAL*, 30-37.

Holambe, A. N., Holambe, S. N. and Thool, R. C. (2010). Comparative study of Devanagari handwritten and printed character & numerals recognition using Nearest-Neighbor classifiers. *Proc. 3rd ICCSIT*, 426-430.

Kudo, T. (2005). CRF++: Yet another CRF toolkit. Retrieved from: http://code.google.com/p/crfpp/

Lafferty, J., Mccallum, A. and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequential data. *Proc. 18th ICML*, 282-289.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. on System, Man, and Cybernetics*, 9, 62–66.

Shetty, S., Srinivasan, H. and Srihari, S. (2007). Handwritten word recognition using conditional random fields. *Proc. 9th ICDAR*, 2, 1098 - 1102.

Shoombuatong, W., Traisathit, P., Prasitwattanaseree, S., Tayapiwatana, C., Cutler, R. and Chaijaruwanich, J. (2011). Prediction of the disulphide bonding state of cysteines in proteins using conditional random fields. *International Journal of Data Mining and Bioinformatics*, 5, 449-464.

Subpaiboonkit, S., Thammarongtham, C. and Chaijaruwanich, J. (2012). RNA Family Classification Using the Conditional Random Fields Model. *Chiang Mai Journal of Science*, 39.

Sutton, C. and Mccallum, A. (2007). *Introduction to Statistical Relational Learning*. MIT Press.

Zhou, X. D., Liu, C. L. and Nakagawa, M. (2009). Online Handwritten Japanese Character String Recognition Using Conditional Random Fields. *Proc. 10th ICDAR*, 521-525.