Development of the application to convert images into Thai characters

Weerachai Jonburom[1, a] and Niwat Angkawisittpan[2, b]
[1] Department of Electrical Engineering, Faculty of Industry and Technology,
Rajamangala University of Technology Isan Sakon Nakhon Campus,
199 Village No. 3, Phungkon, Sakon Nakhon, Thailand
[2] Research Unit for Computational Electromagnetics and Optical Systems (CEMOS)
Faculty of Engineering, Mahasarakham University, Kantarawichai, Maha Sarakham, Thailand.

[a] weerachai.jo@rmuti.ac.th, [b] niwat.a@msu.ac.th

Abstract

This paper presents the application development for converting images to thai characters. The use of the artificial neural network in ThaiOCR applications on the android operating system can dramatically simplify the code and improve the quality of recognition while achieving good performance. Another benefit of using a neural network in ThaiOCR is the extensibility of the system's ability to recognize more thai character sets than initially defined. Experimental results show that at the leaning rate น ด 8000 epochs the obtained RMS error value is 0.00091. However there are some confused thai characters that occured on dissimilar fonts. This drawback can be solved with some additional preprocessing methods such as normalization, thinning and recognition.

Keywords: *Thai optical character recognition*

1. Introduction

There are many different approaches to thai optical character recognition[1] (ThaiOCR) problem. One of the most common and popular approaches is based on neural networks, which can be applied to different tasks, such as pattern recognition, time series prediction, function approximation, clustering, etc.

The use of the artificial neural network[2] in ThaiOCR applications on android operating system can dramatically simplify the code and improve the quality of recognition while achieving good performance. Another benefit of using a neural network in ThaiOCR is the extensibility of the system able to images recognize[3] more thai character sets than initially defined.

Basically, an artificial neural network is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. Artificial neural network (ANN) is also a highly parallel system that processes information through many interconnected neurons that respond to inputs through modifiable weights, thresholds, and mathematical transfer functions.

2. Operating Principle

We composed a feedforward neural network (FFNN) according to and used principles of for word segmentation.

Algorithm Structure

Our implementation and mainly most of the ThaiOCR applications have a process flow like below.

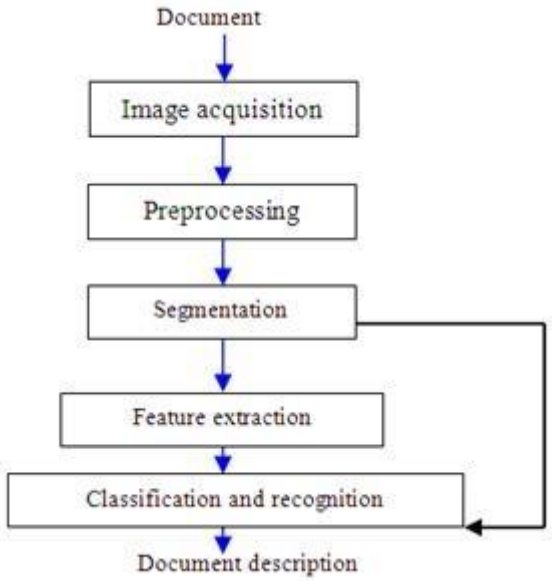

Figure 1 Algorithm Structure

Image Acquisition

In our project, to train the network a computer generated images of alphabet thai characters are used. Developed software automatically generates the images of the thai characters with a specific font and stores the information of these images to use later for the training of the network.

The trained thai characters are as follows;

'0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '๐', '๑', '๒', '๓', '๔', '๕', '๖', '๗', '๘', '๙', 'ก', 'ข', 'ค', 'ฆ', 'ง', 'จ', 'ฉ', 'ช',

'ซ', 'ฌ', 'ญ', 'ฎ', 'ฏ', 'ฐ', 'ฑ', 'ฒ', 'ณ', 'ด', 'ต', 'ถ', 'ท', 'ธ', 'ฝ', 'พ', 'ฟ', 'ภ', 'ม', 'ย', 'ร', 'ล', 'ว', 'ศ', 'ษ', 'ส', 'ห', 'ฬ',

'อ', 'ฮ', '-ะ', '-า', '-ิ', '-ี', '-ึ', '-ื', '-ุ', 'เ-ะ', 'เ-', 'แ-ะ', 'แ', 'ฤ', 'ฤา', '-ำ', 'ไ', 'ใ', 'ฯลฯ', 'ๆ', '-่', '-้', '-๊', '-๋', '-์', '-็'

In the application, there is another part that allows us to test the trained network with a single thai character after the training period. In this section, the network can be tested with different thai fonts than training font. By using this section, an attitude of the network with different fonts can be tested separately.

Also a simple application was developed to test the segmentation part of the system. In this section recognition of the images which include just one line of string which consists of the recognized thai characters.

Preprocessing

The ThaiOCR systems run some code to enhance and improve the quality and readability of the image such as noise removal, normalization, training, recognition, etc. In this project, we just implemented the grayscale conversion for the images. Implementation of some of the mentioned methods will also definitely improve the performance and high accuracy of the current developed system.

Segmentation

The image segmentation[4] of the characters the principles in the study were implemented. This method relies on the horizontal and vertical projection of the digital image. The image processing and digit classifiers algorithm[5,6] simply check the valleys in the projections.

Figure 2 Digital Image

In some cases it is not enough to decide segmentation of thai characters with these valleys. We also developed a tracking algorithm which checks if there was a clear path to upside of the string.

Figure 3 Clear path to upside of the string

Feature Extraction

After segmentation, the image of each character in the whole image is ready to be prepared as a network input. So feature extraction means, preparing an organized data about the image of each character for training the network. There are different approaches for this purpose. For example, some studies prefer to use moments or centers of gravities of the image fragments after dividing images into a couple of fragments. We have chosen the approach suggested in the study.

For every thai character, a 10 x 15 (height x width) matrix is generated. And values in this matrix are scaled to [0, 1]. After this process, this 10x15 matrix is going to be the input for the network.

Figure 4 Feature Extraction

FFNN Structure and Recognition

This step, we use the MATLAB[7] program for train feed forward neural network (FFNN) with a sigmoid activation function that has three layers was build like in the study. Some neurons in the input layer, hidden layer, and output layer are (10x15) 150 (matrix size), 50 and 36 (num of characters) respectively. Network structure is as follows;

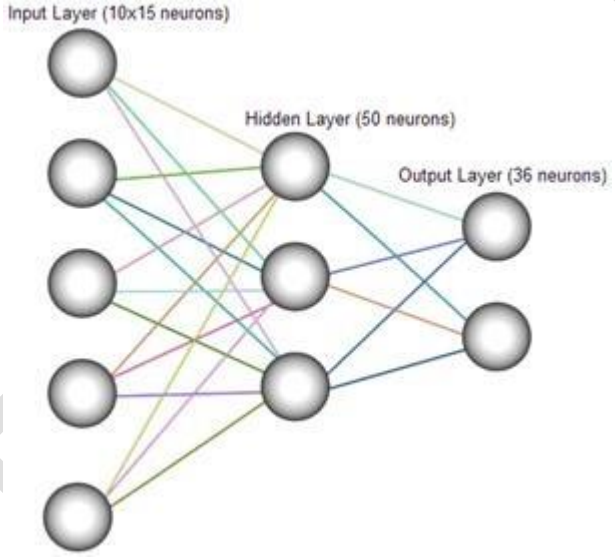

Figure 5 Feedforward neural network (FFNN)

Network input are grayscale values scaled to [0, 1]. And inputs values are “0” and “1” values organized like below;

Output Layer Neurons	Character
1000000000000000000000000000000000000000	0
0100000000000000000000000000000000000000	1
0010000000000000000000000000000000000000	2
0001000000000000000000000000000000000000	3
0000100000000000000000000000000000000000	4
0000010000000000000000000000000000000000	5
0000001000000000000000000000000000000000	6
.....	
0000000000000000000000000000000000000001	ก

Figure 6 Network input values

During the tests, the network which is trained with a learning rate of 0.9 and the number of epochs of 8000 gave good results.

Typical error change during training for first 300 epochs is like below;

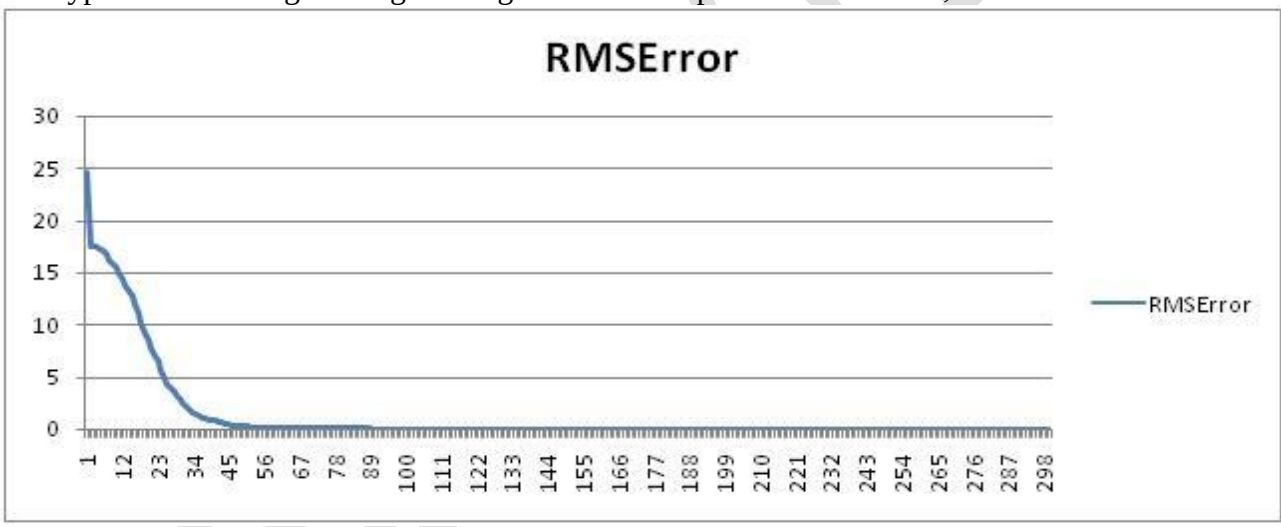

Figure 7 During the tests

After 8000 epochs the obtained RMS error value is 0.00091.

Training and testing with dissimilar fonts may cause some confused thai character pairs. For example, on testing the network with the “Courier New” font, after training with the “Tahoma” font (dissimilar fonts) the confused thai character pairs below emerged.

Courier New(Test Font)	ก	ข	ค	ง	จ	ช	ซ
Tahoma (Training Font)	ก	ข	ค	ง	จ	ช	ซ

Figure 8 Training and testing font

But system can successfully recognize similar fonts.

Developed application

An application is developed for the software project. It simply has three parts like mentioned before.

1. In training, part software generates automatically images for the characters that will be trained based on selected font. In the same section, desired values for some epochs and learning rates can be chosen. Also the status of iteration and RMS error and can be viewed.

Figure 9 Training font

2. After training the network, tests can be applied with different fonts for the trained network.

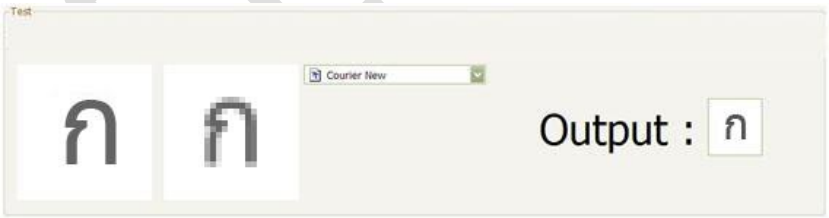

Figure 10 Different fonts for the trained network

3. The application part, images with one line of string can be tested.

Figure 11 Testing application

Conclusion

The developed system has high accuracy on similar fonts at 8000 epochs the obtained RMS error value is 0.00091. On dissimilar fonts, some confused characters occurred. This drawback can be solved with some additional preprocessing methods such as normalization, thinning and recognition.

And also all our experiments were performed using the ideal training set. Possible future research can be done with the implementation of filtering reduction algorithms to get good results with noisy data. It also should be considered that in real-world data there are always orientation errors. So the algorithm should calculate the angle and rotate the lines to obtain a straight data set.

Acknowledgements

The research uses an imaged processing laboratory supported by the Faculty of Industry and Technology, Ragamanggala University of Technology Isan, Thailand.

References

[1] R SANJEEV KUNTE. (2006). **A simple and efficient optical character recognition system**

for basic symbols in printed Kannada text. Department of Electronics and Communication. S

J College of Engineering.

[2] Ali K. Al-Omari Saleh. (2009). Digital Recognition using Neural Network. **Journal of**

Computer Science. Vol. 5. Issue 6 : 427-434.

[3] Kussul, E., T. Baidyk, D. Wunsch, O. Makeyev and A. Martin, 2006. **Image recognition systems based on random local descritors**. Proceeding of the International Joint Conference on Neural Networks, July 16-21, IEEE Xplore Press, USA., pp: 2415-2420. DOI: 10.1109/IJCNN.2006.247067

[4] Westall, J.M. and M.S. Narasimhsa, 1997. **An evolutionary approach to the use of neural networks in the segmentation of handwritten numerals**. Int. J. [4]Patt. Recog. Artifi. Intell. 11: 717-735. http://cat.inist.fr/?aModele=afficheN&cpsidt=2815709

[5] Grother, P.J. and G.T. Candela, 1993. **Comparison of handprinted digit classifiers**. Technical Report NISTIR 5209, NIST. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.8934

[6] Rafael C. Gonzalez and Richard E. Woods, 2002. **Digital Image Processing**. 2nd Edn., Prentice-Hall, Inc., USA., ISBN: 10: 0201180758, pp: 28-29.

[7] Rafael Gonzalez, C., E. Richard Woods and L. Steven Eddins, 2003. **Digital Image Processing using MATLAB**. 2nd Edn., Prentice Hall, USA., ISBN: 10: 0130085197, pp: 624.

Bibliography

Mr.Weerachai Jonburom
Department of Electrical Engineering, Faculty of Industry and Technology,
Rajamangala University of Technology Isan Sakon Nakhon Campus,
199 Village No. 3, Phungkon, Sakon Nakhon, Thailand.
Email: weerachai.jo@rmuti.ac.th

Assoc.Prof. Niwat Angkawisittpan
Research Unit for Computational Electromagnetics and Optical Systems (CEMOS), Faculty of Engineering, Mahasarakham University,
Kantarawichai, Maha Sarakham, Thailand.
Email: niwat.a@msu.ac.th