National Electronic and Computer
Technology Center, Image Technology Lab
National Authority for Sciences and Technology,
Information Technology Research Institute

Lao Optical Character Recognition (Lao OCR)

Mr. Wasin Sinthupinyo, Acting Chief of Image Technology Lab
Ms. Chittaphone Chansylilath, Researcher of Development
And Research Training Center

1. Abstract

Optical Character Recognition (OCR) system is an essential device in converting information available in printed format to electronic format. The main advantages of automating this process by a computer based application, i.e. OCR are fast, economical and not labor intensive. In the localization point of view, developing an OCR system for local language scripts enables to generate huge amount of local language content which can be circulated among the local community over the World Wide Web and any other electronic medium.

Under the MOU in collaboration on research and Development of Electronics, Computer, Telecommunication and Information Technology; between Information technology Research Institute, National Authority for Science Technology (NAST), Lao PDR and National Electronics and Computer Technology Center, National Science and Technology Development Agency, Thailand.

The collaboration is including developing Lao Optical Character Recognize (OCR) application and Lao Text-to-speech (TTS). The developing program was designed in full time and on job training a Lao researcher in NECTEC for Lao script base on the experience acquires in the research and developing process of Thai OCR and Thai TTS.

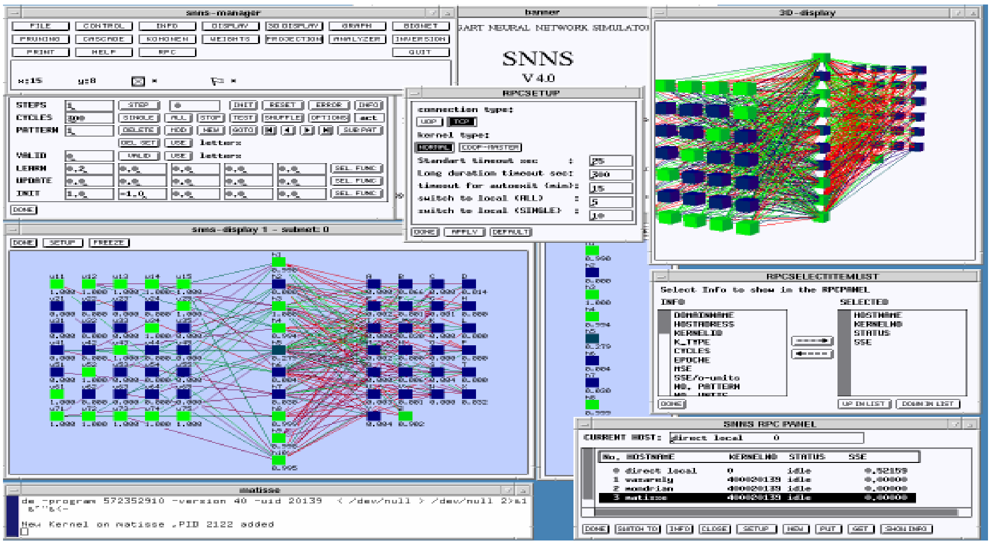
This report describes the process of developing Lao OCR in details, Presents an analysis of Lao writing system and characters set for OCR including training, Implementing and the result of evaluation, also the error and problem of OCR system and future plan to improve.

2. Introduction

SNNS is stand for: Stuttgart Neural Network Simulator is a software simulator for neural networks on Unix workstations developed at the Institute for Parallel and Distributed High Performance Systems (IPVR) at the University of Stuttgart. The goal of the SNNS project is to create an efficient and flexible simulation environment for research on and application of neural nets [1].

The SNNS simulator consists of four main components:

- Simulator kernel
- graphical user interface
- batch simulator version snnsbat, and
- network compiler snns2c



```
connection type:  UDP  TCP
kernel type:  NORMAL  COOP-MASTER
Standart timeout sec         :  25
Long duration timeout sec:  300
timeout for autoexit (min):  15
switch to local (ALL)        :  5
switch to local (SINGLE)   :  10
```

```
CURRENT HOST: direct local      0
No. HOSTNAME           KERNELNO   STATUS    SSE
  0 direct local       0          idle      0.52159
  1 vasarely           400020139  idle      0.00000
  2 mondrian           400020139  idle      0.00000
  3 matisse            400020139  idle      0.00000
```

```
de -program 572352910 -version 40 -uid 20139  < /dev/null > /dev/null 2>&1
 &"'&<-

New Kernel on matisse ,PID 2122 added
```

(Figure 1: GUI of SNNS v4.1 *http://www.ra.cs.uni-tuebingen.de/downloads/SNNS*)

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well [2].

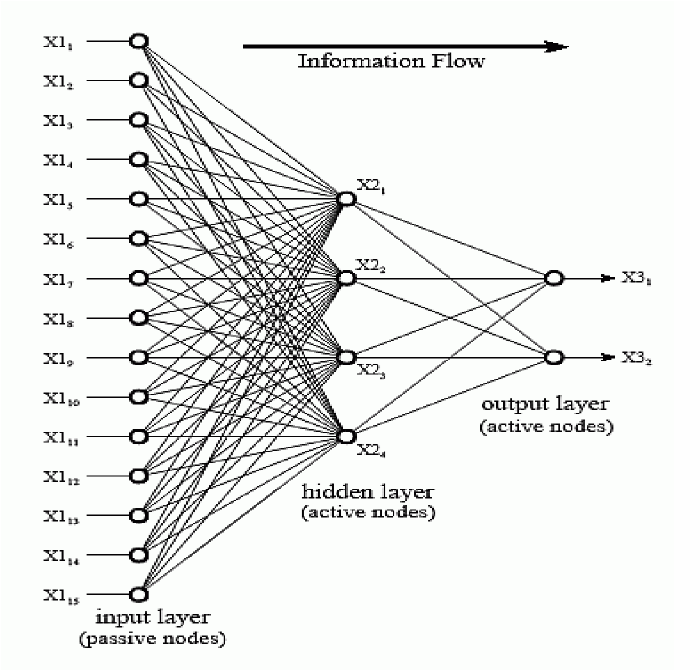
2.1 Architecture of neural networks

2.1.1 Feed-forward networks

Feed-forward ANNs (figure 1) allow signals to travel one way only; from input to output. There is no feedback (loops) i.e. the output of any layer does not affect that same layer. Feed-forward ANNs tend to be straight forward networks that associate inputs with outputs. They are extensively used in pattern recognition. This type of organization is also referred to as bottom-up or top-down [2].

2.1.2 Feedback networks

Feedback networks (figure 1) can have signals travelling in both directions by introducing loops in the network. Feedback networks are very powerful and can get extremely complicated. Feedback networks are dynamic; their 'state' is changing continuously until they reach an equilibrium point. They remain at the equilibrium point until the input changes and a new equilibrium needs to be found. Feedback architectures are also referred to as interactive or recurrent, although the latter term is often used to denote feedback connections in single-layer organizations [2].



(Figure 2: An example of a simple feedforward network; http://www.emsl.pnl.gov:2080/docs/cie/neural/neural.homepage.html)

3. Analysis of Lao Writing System

3.1. Lao Script

Lao scripts originated from the writing system used by Thais in the 13th century which was in turn a modified version of the Khmer scripts that was adapted by the Cambodian scholars from Sanskrit *Devanagari* scripts. With the introduction of new spelling system in 1975 Lao writing system was completely standardized and simplified by reducing its alphabet only to letters actually pronounced [3].

3.2. Structural Features

Lao letters are written from left to right and vowel modifiers can occur before, after, above and below the base consonants. The tone marks occur only on the top of base consonants; if there is a vowel modifier top of the base consonant the tone mark is written above the vowel modifier (see Figure 2.2.1).

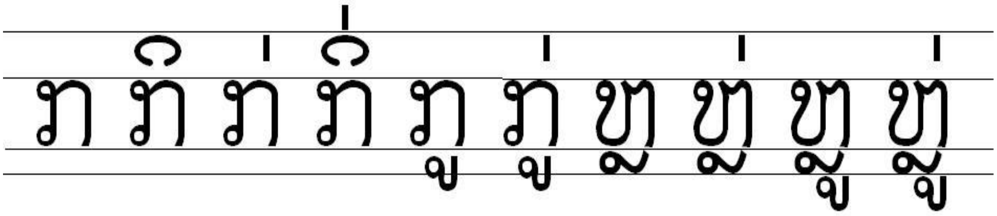


Figure 3: Typical Layers in Lao Writing System

As Figure 3 depicts, there can be two levels above the main base consonant and normally there can be only one line below the main base consonant except for the character □□ as the character itself has a disconnected component below it as part of the consonant (see the last four characters in Figure 3).

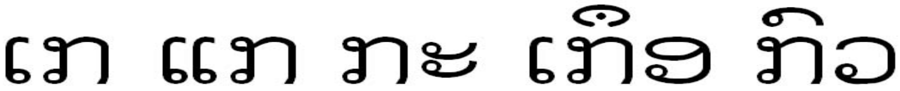


Figure 4: Vowel Modifiers before and after base characters

Figure 4 shows how vowel modifiers can be combined with base characters in Lao script. As shown in the first two characters, in some cases same symbol is repeated to denote another vowel modifier.

4. Training data

According to Lao fonts is not so difference shape, so we consider Lao language as 1 group as a training data

a. Training data preparation

- **Font Selection**

10 common fonts have been used in Lao OCR, in the list bellow are font name setting (setting the name of font by using 2 characters as the font name)

FONT NAME	NAME USED
Alice0_2000	A0
Alice1_2000	A1
Alice0 Lao	AL
Chantabouli95	C9
ChantabouliLao	CL
FontLao1 TT	FL
Saysettha OT	SO
Saysettha2000	S2
Saysettha 95	S9
Saysettha Lao	SL

(Table 1: Name list for Lao COR)

- **Character Set for OCR**

The character set are consonant characters, vowel characters, tone marks, Lao digits and special symbols. It contains many combined glyph characters. The mixing with alphanumeric characters can be handled only by switching the font list. But all of this is not enough for OCR system, because the main problem or main error in OCR is the characters which touch with each other or connected characters, therefore we was considered for touching characters and seem to be big characters group as in the list below:

Lao Character set	Name Used
Lao Consonant	LC
Lao Vowel	LV
Lao Number	LN
English Capital Letter	EC
English Small Letter	ES
English Number	EN
Special Character	S1
Touching Character 1	T1
Touching Character 2	T2
Touching Character 3	T3
Touching Character 4	T4
Touching Character 5	T5
Touching Character 6	T6

(Table 2: Character set for Lao COR)

- **Type of character**

Prepared for 2 type of character format: normal and bold

- Normal-> N
- Bold -> B

- **Resolution and scanning time**

Each page of training data will be scanned in 3 resolutions: 200 dpi, 300 dpi and 400 dpi.

- 200 dpi 2 times -> 21 22
- 300 dpi 5 times -> 31 32 33 34 35
- 400 dpi 2 times -> 41 42

- **Size of Characters**

10 12 14 16 18 20,

So the total number of each character is: 10 x 2 x 9 x 6 = 1,080 images

b. Training process

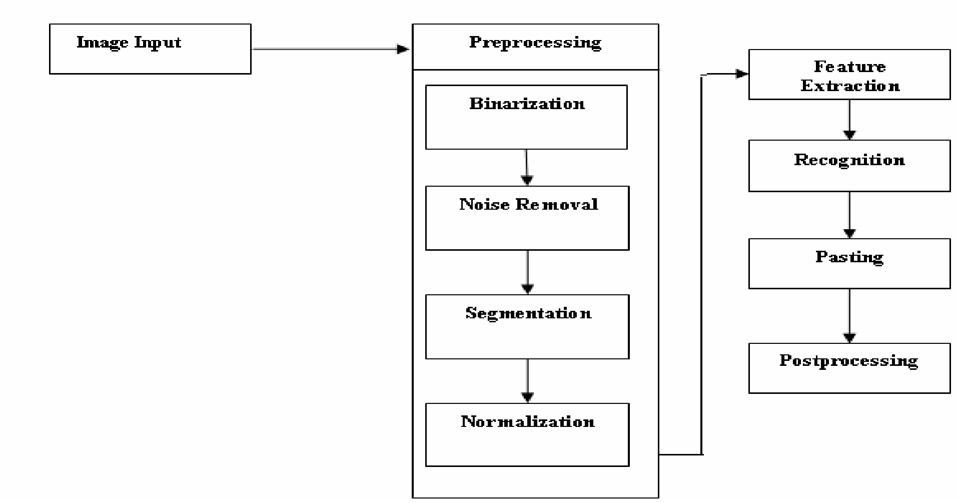
- Scan the training data
- Crop and Rotate the training image after scanning
- Remove noise
- Segments into character and save it in specific name format for example: A0LCN21_10_161_1.bmp is a Alice0 Lao font: A0, character set is Lao Consonant: LC, Normal Type: N, Resolution 200 dpi, first scanning times, Size 10 and ASCII code: 161 respectively

5. Design and Implementation

The Lao OCR application was implemented by neural network for character recognition and Borland C++ builder as the programming language and application creator. Lao OCR application are accepts any standard format as its input images and generates a unicode text as output file. The OCR consists of modules which will be describes in the next topic of this report.

a. Overview of the OCR

The main modules of OCR are shown in Figure 3 and for detail description of each module is given in Section 4.2.



(Figure 5: Overview of the OCR)

6.1.1. Preprocessing

We divided Preprocessing module in to four subtasks:

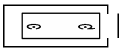
- **Binarization**

Reading input image format (bmp, jpeg, gif) and convert into DIB (Device-Independent Bitmap Graphic) specific, then convert into black and write image, 0 and 1 according to the scanned values of image. The background is completely white and foreground is completely black.

- **Noise Removal**

  This procedure removes noise exist in the input image that may degrade the performance of the system recognition process. Currently this module removes only isolated noise that interrupt the smooth running of the system.

- **Segmentation**

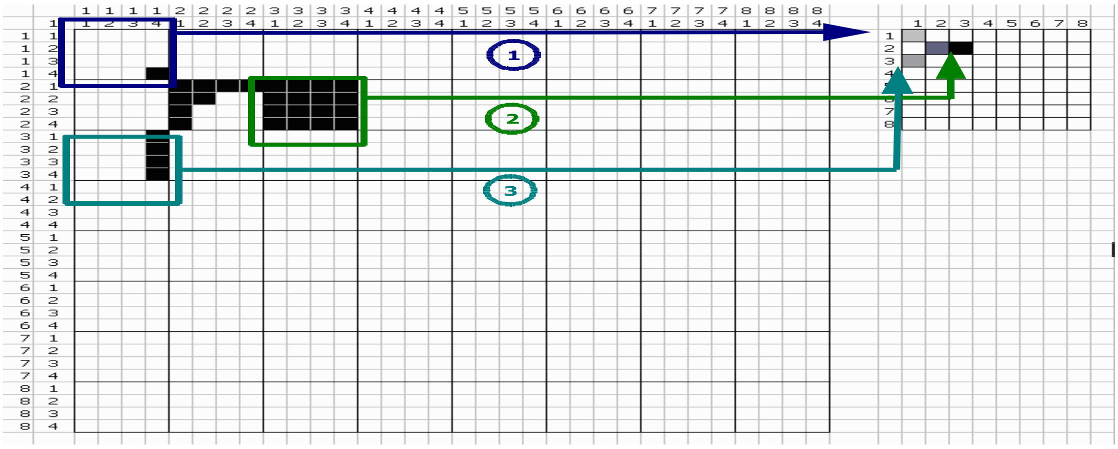
  Fast Connected Component Labeling Algorithm Using a Device and Conquer Technique by Jung-Me Park, Carl G. Looney and Hui-Chuan Chen from CATA 2000 conference on computers and their application, PP 373-376 Rec.2000. Used to get the connected character and segmented it into characters, for Lao Characters is still have some problem with some characters and can not segment correctly like:  is segmented in to “black dot (.)” and “ີ or ິ”, so right now we just consider only “ີ” and “ິ” and other one in English is ‘i’ and ‘j’ we also have to remove the top part and all of this can be solve in pasting module.

- **Normalization**

  Each segmented character is scaled into 32 X 32 and normal into 8 X 8 as the features of each character which enables the recognition algorithm to compare them with the output of SNNS training data.

6.1.2 Feature Extraction

Get any images from user and scaled into 32 x 32 pixel, then take 4x4 pixel as a special features, so each character will have 8x8 = 64 features(see the picture bellow) and include the ratio of width per height and total features 65 features.



(Figure 6: table with pixels comparison)

- in 4 x 4 there is 1 black point, it mean that in table 8x8 in point (1,1) will have the weight of black pixel: 1/16 = 0.0625
- in 4 x 4 there is 16 black point, it mean that in table 8x8 in point (3,2) will have the weight of black pixel: 16/16 = 1.000
- in 4 x 4 there is 4 black point, it mean that in table 8x8 in point (1,3) will have the weight of black pixel: 1/16 = 0.250

6.1.2 Recognition

Using Neural Networks Method and training tool is SNNS v 4.1. We assigned all Lao characters (Consonants, Vowels, Number, English small and capital letter, special characters, English number and touching characters) into 1 group of training set and used 65 of input Nodes, 65 hidden nodes and 260 output nodes for training. The recognize engine will uses weighted that we trained and Compute of recognizes.

6.1.3 Pasting

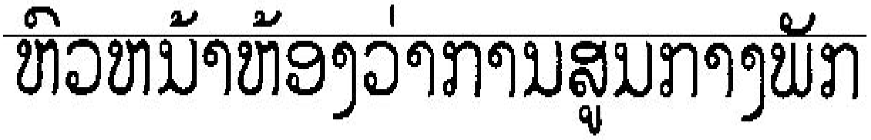
It is one important module in Lao OCR system; main works of this module is to map the result of Recognition module with Lao Unicode and paste it into text editor and generate the output text file.

6.1.4 Post processing

As Lao has a complex writing system with many levels of vowel modifiers and tone marks Post processing step plays an important role in the OCR system and considerable efforts have been made to handle the issues arise in combining the modifiers to the base consonant.

a. Combining Lines

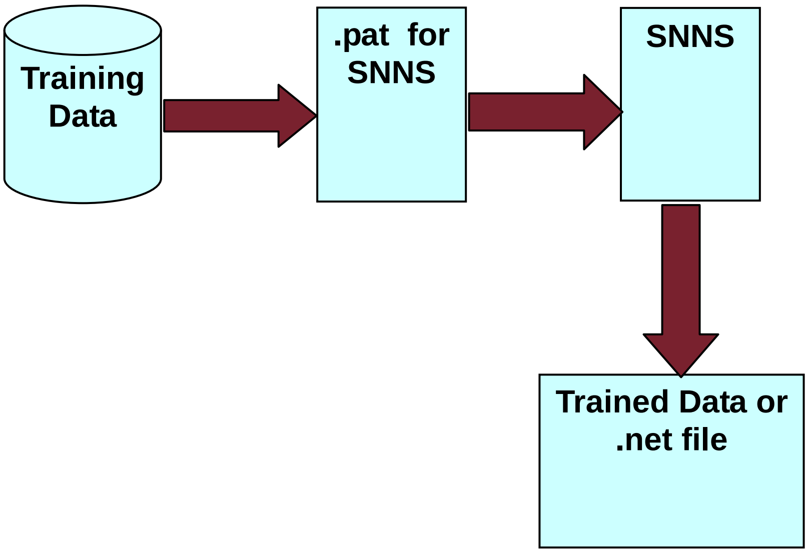
In some cases vowel modifiers and tone marks form separate lines as shown in Figure 3 in which case it necessary to combine those lines into the main consonant line.


ຫົວຫນ້າຫ້ອງວ່າການສູນກາງພັກ

(Figure 7: Top vowel modifiers form a separate line)

According to the result from recognize and paste modules, is not 100% accuracy. To get more accuracy; we can improve the result by text processing such as: Dictionary, word list or spell checker and some other tool, bri-gram, HMM – base and so on

6.1.5 SNNS training process

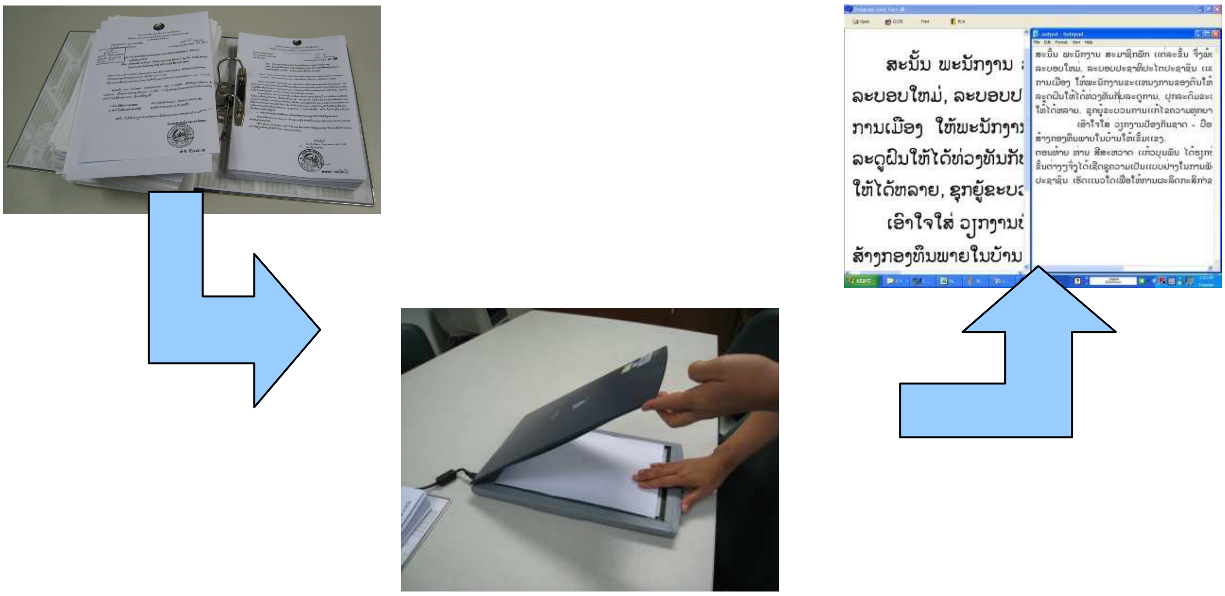


(Figure 8: SNNS Training process)

- Making the table to map between ASCII and node of Neural Network
- Develop software to read all characters image and generated text file as a pattern file to load SNNS for SNNS Training names: PrepareSNNS Modules.
- Train Neural Network by using SNNS on Linux
- Upload Train.pat
- Cut Code enter by using dos2unix
- Call SNNS
- Generated .net or network file choose tap Sheet : Bignet choose input node : 65 then Insert press type to change type hidden node : 65 then Insert, press type change type out node : 260, press Full connected and Create Network
- Save and load patter file choose Tab Sheet : File choose net to save file net and press pat then choose pat file to load pattern into Memory

- Train choose Tab Sheet : Control choose Learning func : Std_Backprogation Update func: Topological order Init func : Randomize Weights press Init and Shuffle set Learn is 0.2 02 set Cycle by the number of Cycle which is wants to press, press All to Train, when finished training press valid to view the result of the training
- Save net to use in OCR application
- Develop OCR application, by using the SNNS Trained data for Characters recognition

7. Process of using Lao OCR



(Figure 9: Use of Lao OCR application)

8. Evaluation & Results

8.1 Evacuations for official documents only

Evaluate with Lao official document and removed some unwanted such as big noise, image and some the text with underline or italic (for compare with old Lao OCR systems)

File Name	Total Characters	Old Lao OCR system (Before develop in NECTEC)		New Lao OCR system (Developed in NECTEC)	
		Error Number	% Correct output	Error Number	% Correct output
13_3_300.bmp	2155	450	79.12	98	95.45
BiglaotextCon.bmp	1086	200	81.58	10	99.08
TextExample1.bmp	1778	321	81.95	27	98.48
TextExample2.bmp	1703	344	79.8	34	98
TextExample3.bmp	1710	315	81.58	30	98.25
TextExample4.bmp	1680	103	93.87	29	98.27
Testocr001.bmp	654	45	93.12	13	98.01
5_2_300.bmp	1660	220	86.75	38	97.71
3_1_300.bmp	1146	250	78.18	35	96.95

Average	83.99	Average	97.8

(Table 2: Character set for Lao COR)

8.2 Evaluation for all kind of documents such as: Book, Magazine and some newspaper (see the detail of evaluation in the table below)

File Name	Total Characters	Incorrect Character	Error Percentage %	Character Confusing
Book001	2074.00	290	14	ນ້ຳ ເປັນ ປູ່ ◌ັາ
Book002	1133.00	459	41	◌່ ມັກບໍ່ຄ່ອຍເຫັນ
Book003	1005.00	251	25	ເມື່ອຢູ່ເທິງ ◌ິ ◌ີ ◌ຶ ◌ື
Book004	941.00	406	43	***
Book005	1798.00	202	11	***
Book006	1565.00	439	28	***
Book007	1406.00	227	16	***
Book008	1598.00	206	13	***
Book009	1400.00	181	13	Xາ ເປັນ ຊຸ
Book010	1359.00	211	16	ຫ ເປັນ ທ
Book011	1169.00	303	26	ຫ ເປັນ ທ
Book012	1518.00	209	14	Xາ ເປັນ ຊຸ
Book013	1044.00	108	10	Xາ ເປັນ ຊຸ
Book014	1082.00	343	32	ໜ ເປັນ ໝ
Book015	1644.00	225	14	Xາ ເປັນ ຊຸ
Book016	1438.00	124	9	ໜ ເປັນ ໝ
Book017	1474.00	183	12	ນ ເປັນ ໜ
Book018	1101.00	150	14	◌່ ມັກພົບເຫັນຕາມຄຳສັບຕ່າງໆ
Book019	1189.00	179	15	ຫ ເປັນ ທ
Book020	1132.00	102	9	Xາ ເປັນ ນ,ຊຸ
Gvlt001	1314.00	611	46	ຂ ເປັນ ກ
Gvlt002	900.00	35	4	***

Gvlt003	1609.00	69	4	ໜ ເປັນ ໝ
Gvlt004	1469.00	135	9	ໜ ເປັນ ໝ
Gvlt005	1807.00	97	5	ໜ ເປັນ ໝ
Gvlt006	2042.00	107	5	◌່ ມັກພົບເຫັນໃນທ້າຍຄຳສັບ.ຫຼ ເປັນ 7
Gvlt007	2001.00	63	3	ໜ ເປັນ ໝ
Gvlt008	2073.00	122	6	◌່,◌້ ມັກບໍ່ເຫັນໃນຄຳສັບ
Gvlt009	2459.00	100	4	◌່ ມັກພົບເຫັນຕາມຄຳສັບຕ່າງໆ
Gvlt010	1219.00	191	16	ຣ ເປັນ ຣ
Gvlt011	1910.00	72	4	***
Gvlt012	2134.00	90	4	ໜ ເປັນ ໝ
Gvlt013	2155.00	119	6	ຄຳສັບຕົວໃຫຍ່ເປັນຄຳສັບຕົວນ້ອຍເຊັ່ນ:N ເປັນ n etc..
Gvlt014	2223.00	111	5	◌່ ມັກພົບເຫັນຕາມຄຳສັບຕ່າງໆ
Gvlt015	2230.00	115	5	% ເປັນ o/o
Gvlt016	2145.00	92	4	. ມັກມີຕາມໜ້າຄຳສັບ
Gvlt017	1936.00	65	3	ກ ເປັນ n
Gvlt018	1890.00	65	3	. ມັກມີຕາມໜ້າຄຳສັບ
Gvlt019	1710.00	48	3	◌່ ມັກພົບເຫັນຕາມຄຳສັບຕ່າງໆ
Gvlt020	2078.00	69	3	ໜ ເປັນ ໝ
Gvlt021	1817.00	76	4	ໝ ເປັນ ໜ
Gvlt022	1006.00	57	6	ໜ ເປັນ ໝ
Gvlt023	541.00	27	5	ໜ ເປັນ ໝ
Gvlt024	355.00	72	20	ຫ ເປັນ ທ
Gvlt025	1273.00	37	3	ໜ ເປັນ ໝ
Gvlt026	257.00	29	11	L ເປັນ i
Gvlt027	1327.00	32	2	ໜ ເປັນ ໝ
Gvlt028	1324.00	62	5	No comment
Gvlt029	786.00	676	86	◌ິ ເປັນ ◌ີ
Gvlt030	992.00	638	64	ຫຼ ເປັນ 7,p
Gvlt031	978.00	840	86	◌່ ເປັນ ,

Gvlt032	938.00	32	3	◌່ ມັກບໍ່ຢູ່ໃນຄຳສັບ
Gvlt033	855.00	27	3	No comment
Gvlt034	803.00	31	4	No comment
Gvlt035	637.00	184	29	No comment
Gvlt036	1480.00	65	4	ນ ເປັນ ມ
Gvlt037	1178.00	38	3	◌ິ ເປັນ ◌ີ
Gvlt038	1740.00	79	5	ໜ ເປັນ ໝ
Gvlt039	1426.00	70	5	No comment
Gvlt040	876.00	60	7	ເX ເປັນ (
Maga001	2608.00	838	32	ນ ເປັນ ໜ
Maga002	777.00	57	7	Xໆ ເປັນ ຊ
Maga003	1489.00	271	18	Xໆ ເປັນ ຊ
Maga004	986.00	197	20	ໜ ເປັນ ໝ
Maga005	2168.00	787	36	Xໆ ເປັນ 1,6,9
Maga006	2596.00	598	23	◌ັ ເປັນ◌້
Maga007	2343.00	568	24	ນ ເປັນ V
Maga008	3206.00	611	19	ຫ ເປັນ ທ
Maga009	1534.00	302	20	ທ ເປັນ ຫ
Maga010	1380.00	226	16	ຫ ເປັນ ທ
Maga011	1068.00	241	23	ເX ເປັນ ໆ
Maga012	779.00	57	7	◌່ ເປັນ◌້
Maga013	1538.00	237	15	ບ ເປັນ ພ
Maga014	657.00	65	10	◌ໍ ເປັນ .
Maga015	665.00	37	6	Xໆ ເປັນ 9
Maga016	769.00	57	7	ນ ເປັນ ກ
Maga017	1273.00	160	13	Xໆ ເປັນ 9
Maga018	995.00	96	10	◌່ ມັກພົບເຫັນໃນທ້າຍຄຳສັບ, ທ ເປັນ ຫ
Maga019	1247.00	83	7	ນ້ຳ ເປັນ ນຊ
Maga020	1640.00	168	10	Xໆ ເປັນ 9

News001	1184.00	780	66	Xໆ ເປັນ ຊຸ
News002	1014.00	427	42	່ ເປັນ ້
News003	916.00	438	48	ຫຼ ເປັນ 7
News004	563.00	182	32	Xໆ ເປັນ S,ຊຸ
News005	288.00	102	35	ໜ ເປັນ ໝ
News006	218.00	201	92	ຸ ມັກບໍ່ພົບເຫັນໃນຄຳສັບເຊັ່ນ: ບຸກ,ສຸ
News007	1499.00	540	36	No comment
News008	438.00	108	25	ລ ເປັນ ວ, ເX ເປັນ (
News009	334.00	114	34	ກ ເປັນ ຕ,ນ
News010	359.00	162	45	ຄ ເປັນ ດ,່ ມັກພົບເຫັນຕາມຄຳສັບຕ່າງໆຢູ່ໜ້າ, ກາງ, ທ້າຍຄຳສັບ
News011	224.00	210	94	Xະ ເປັນ ັັ
News012	387.00	305	79	ກ ເປັນ ນ,ຕ
News013	468.00	126	27	ແລະ ເປັນ ເໜະ
News014	879.00	212	24	ິ ເປັນ ຍ
News015	739.00	203	27	: ເປັນ ..
News016	1162.00	552	48	ໝ ເປັນ ໜ
News017	552.00	185	34	ໍ ເປັນ .ລ ເປັນ ວ;ເX ເປັນ C,(;ແລະ ເປັນ ໜະ;ກ ເປັນ ຕ
News018	446.00	113	25	ກ ເປັນ ຕ,ດ
News019	306.00	69	23	ກ ເປັນ ຕ
News020	530.00	150	28	ກ ເປັນ ຕ,ດ ;ລ ເປັນ ວ;ູ ມັກພົບເຫັນຕາມຄຳສັບ;ກ ເປັນ ຕ;ິ ເປັນ ິ
	129261	21431	17	83

(Table 3: Lao OCR Evaluation 2)

9. Summary

- The accuracy of new Lao OCR systems is about 97.8% with the new document and removed some unwanted and about 83% for books, Magazine and some newspaper
- Time rate is about 10 second /1 page
- Lao OCR application was developed by modifying Thai Source code
- For more accuracy and get more an effective, should add more characters or type of characters such as italic and underline etc.
- Still have some problem with Line Separate (some time wrong)

10. Discussion

The current Lao OCR has been developed by adapting the algorithm used in Thai OCR developed at National Electronic and computer Technology (NECTEC), Thailand. The users of the OCR application can give any standard image format as the input and the OCR application generates its output as a Unicode file. Current OCR recognizes printed Lao characters of ten popular Lao fonts namely, Alice0 Lao, Alice1_2000, Alice2_2000, Chanthabuli Lao, Chanthabuli 95, Saysettha Lao, Saysettha 2000, Saysettha 95 and FontLao1 with an accuracy of ~98.7%. The error caused mainly due to confusion groups and poor segmentation. The OCR has been implemented in Borland c++ builder 6.0 using object oriented programming concepts. The modules can be reused and easily extendable. They can be used in similar applications esp. in OCR applications for other languages.

Reference

[1] http://www.ra.cs.uni-tuebingen.de/SNNS/announce.html

[2] http://www.emsl.pnl.gov:2080/docs/cie/neural/neural.homepage.html

[3] www.laol10n.info.la/