

Khmer OCR Accuracy Measurement for Unicode and Legacy fonts in Tesseract 4.0 with Deep Neural Network



Presented By:

Ms. Kim Sokphyrum

Ms. Suos Samak

Supervisor:

Mr. Javier Sola

Contents

1

Previous Khmer OCR Development

2

Tesseract 4.0 with Deep Neural Network

3

How-to Fine tune OCR engine

4

Accuracy Measurement &
Comparison

5

Conclusion & Future Work

Mr. Vichet Chea. (2015),
Acc: **92.74%** for 6 Khmer Unicode
fonts. (**Tesseract 3**)

Hann Meng et al. (2014),
Research on theory of ANN
only.

Iech SETHA (2012),
Acc: **99%** for Khmer OS Content
(36pt). (**Edge detection &**
template matching)

Mr. En Sovann (2011),
Acc: **96.1%** (improve on work
of Mr. Krui Vanna)

Mr. ING LENG IENG (2009),
Acc: **98.88%** for Limon R1
(22pt). Acc.**92.48%** for Limon S1
(22pt). (**Discrete Cosine**
Transform, HMM)

Chanoeurn CHEY et al. (2003),
Acc: **92%** for 10 Khmer
characters. (**Lagedre /**
Euclidean)

Mr. PHAN Neth (2016),
Acc: **92%** (12pt) for 12 fonts.
(**LSTM base**)

Vandeth Ith (2014),
Acc: **50%** for Limon S1 for various font size.
(**Tesseract 3**)

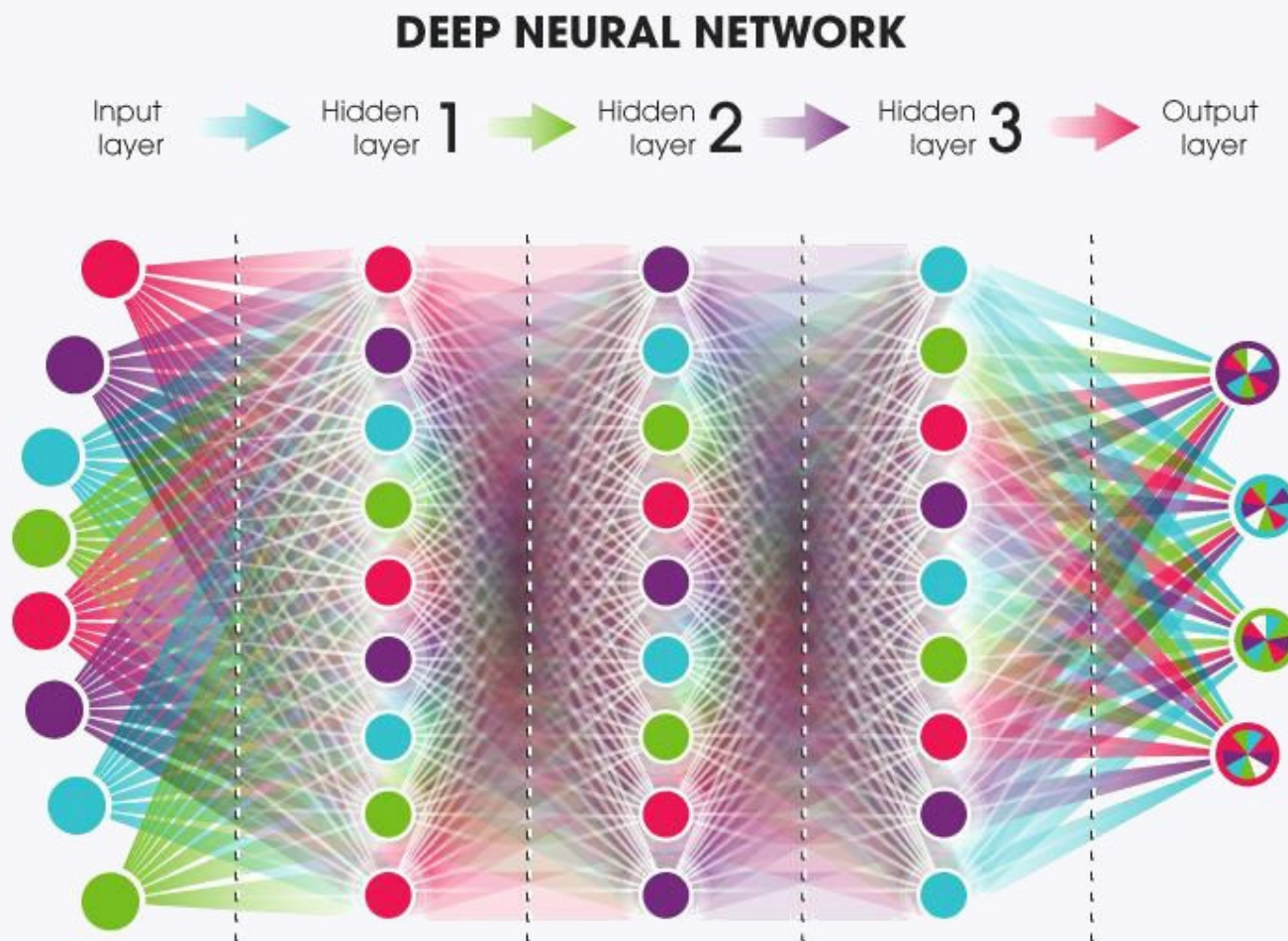
Mr. Danh Hong (2012),
Acc: more than **95%**
(**Tesseract 3**)

Hao Jeudi (2011),
Acc: **96.1%** (improve UI on
work of Mr. En Sovann)

Mr. KRUI VANNA (2011),
Acc: **97.9%** for 19 types of Limon fonts. (**Vertical**
Component extr., Connected Component Analyse,
& Scale Invariant Feature Transform extr.)

Chanoeurn CHEY et al. (2006),
Acc: **92.85%** (22pt), **91.66%** (18pt), &
89.27% (12pt). (**Wavelet Descriptor**)

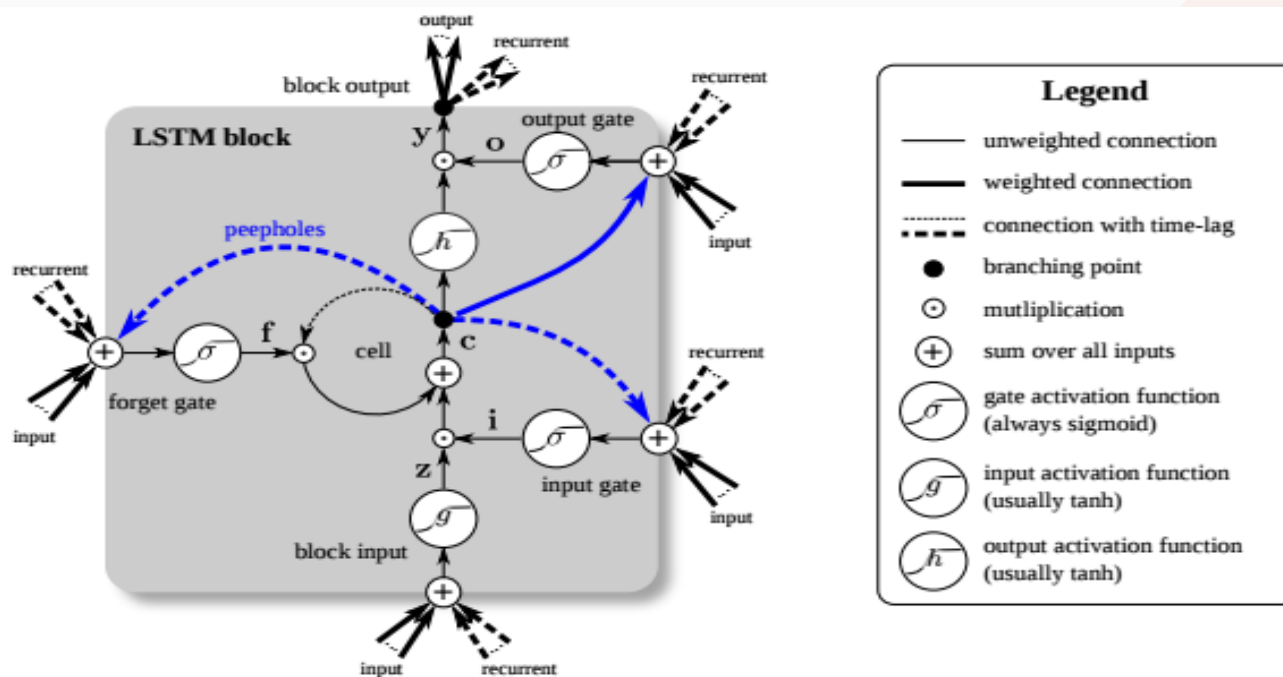
What is Deep Neural Network?



Deep Neural Network = Human Neural Network



- Variable-size Graph Specification Language (VGSL)
- Convolutions & LSTMs



- Variable size images as the input

- Neural Network model, i.e

[1,48,0,1 Ct3,3,16 Mp3,3 Lfys64 Lfx96 Lrx96 Lfx384 O1c1]

[1,36,0,1 Ct3,3,16 Mp3,3 Lfys48 Lfx96 Lrx96 Lfx256 O1c111]

3

How-to fine tune OCR engine

khm langdata

- khm.training_text (**7200** lines)
- khm.numbers (**4526** lines)
- khm.punc (**808** lines)
- khm.training_text_bigram_freqs (**102416** lines)
- khm.training_text_unigram_freqs (**3635** lines)
- khm.word.bigrams (**2461** lines)
- khm.wordlist (**151860** lines)

Total: 272,906 lines

combine_tessdata

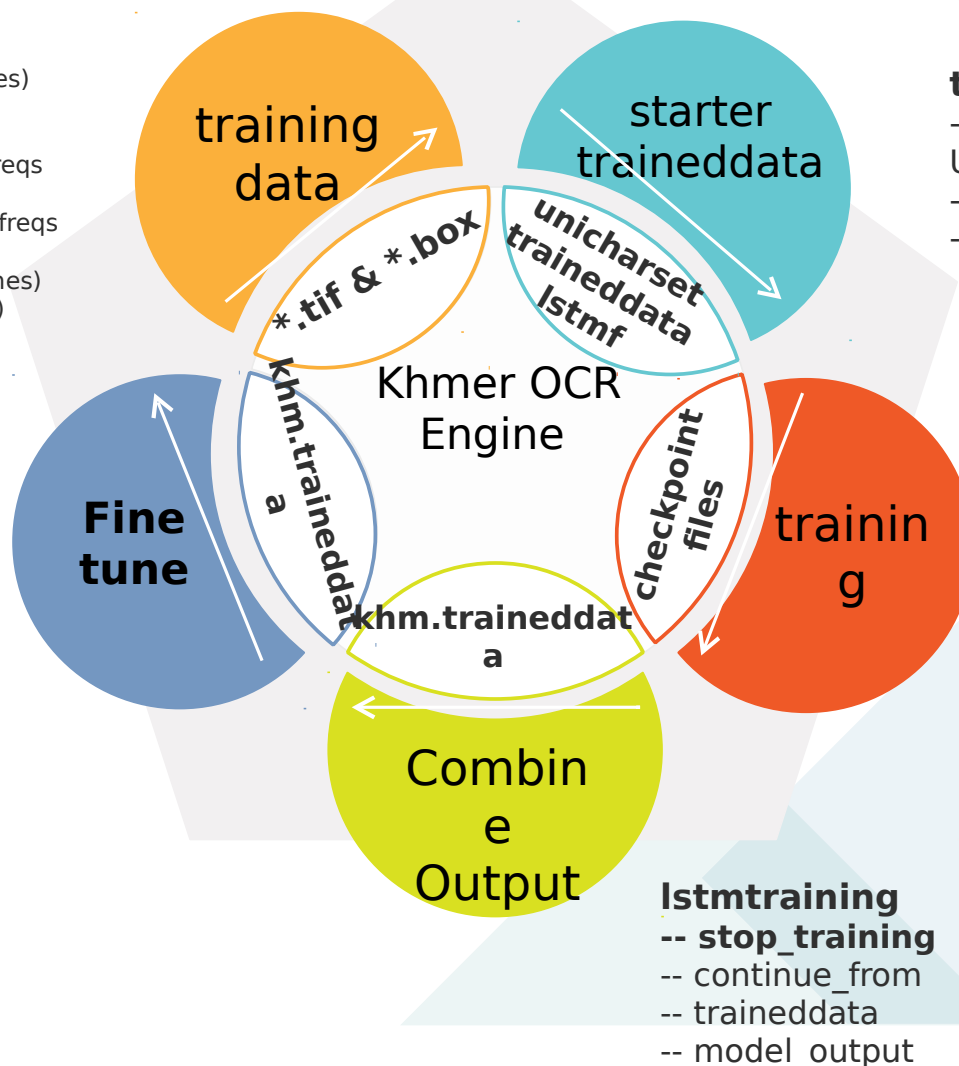
-- e

Istmtraining

- traineddata
- old_traineddata
- target_error_rate
- continue_from

Istmtraining

- **stop_training**
- continue_from
- traineddata
- model_output



tesstrain.sh

```
-- fontlist "LimonS1
Unicode" -- linedata_only
-- lang khm
-- .....
```

Istmtraining

- traineddata
- net_spec
- max_iterations 20000
- model_output

Istmtraining

- **stop_training**
- continue_from
- traineddata
- model_output

How-to fine tune OCR engine (tiff & box files)

មានសេវាករណ៍ស្រស់ស្អាត ព្រោះហេតុនោះ *

1 / 110 4644 125 4681 0
2 > 126 4644 152 4669 0
3 152 4630 171 4689 0
4 မ္ဍ 171 4630 219 4689 0
5 219 4630 236 4689 0
6 ၁ 236 4644 262 4677 0
7 262 4643 278 4677 0
8 လ် 278 4643 340 4676 0
9 လ် 339 4643 384 4676 0
10 မ္ဍ 383 4632 428 4688 0
11 လ် 428 4643 455 4675 0
12 လ် 457 4642 502 4691 0
13 502 4629 516 4691 0
14 မ္ဍ 516 4629 581 4675 0
15 လ် 582 4643 617 4675 0
16 617 4630 632 4686 0
17 ပ္ဍ 632 4630 679 4686 0
18 လ် 681 4642 708 4674 0
19 708 4673 709 4674 0
20 လ် 112 4513 140 4558 0
21 လ် 143 4513 173 4549 0
22 လ် 176 4499 220 4557 0
23 လ် 220 4512 250 4548 0
24 ပ္ဍ 252 4501 321 4557 0
25 လ် 322 4512 352 4548 0
26 မ္ဍ 355 4511 402 4544 0
27 လ် 406 4511 432 4543 0
28 မ္ဍ 435 4511 484 4543 0
29 ပ္ဍ 488 4500 533 4543 0
30 လ် 534 4510 564 4555 0
31 564 4554 565 4555 0
32 မ္ဍ 112 4380 159 4413 0
33 လ် 163 4381 189 4413 0
34 လ် 192 4379 253 4412 0
35 ပ္ဍ 253 4380 287 4412 0
36 လ် 289 4367 323 4424 0
37 လ် 323 4369 382 4425 0

4

OCR Accuracy Measurement Tool

ISRI (Information Science Research Institute) Analytic Tools (Version 6.1)

ISRI Accuracy Commands:

Character Accuracy & Word Accuracy:

- a. Ground truth file**
- a. OCR generated file**
- b. Accuracy report file**

Problem of Word Accuracy for Khmer language:

ISRI toolkit can only produce Cluster Accuracy for Khmer Language.

4

Accuracy Measurement and Comparison (no noise)

Engine: Tesseract 4.0

Net_spec value "[1,48,0,1 Ct3,3,16 Mp3,3 Lfys64 Lfx96 Lrx96 Lfx384 O1c1]"

Testing Text:

Characters :

19618

Clusters :

11272

No.	Font Name	Size	Character Accuracy		Cluster Accuracy	
			Errors	Accuracy	Errors	Accuracy
1	Khmer OS	12	1495	92.38%	650	94.23%
2	Khmer OS Battambang		562	97.14%	140	98.76%
3	Khmer OS Bokor		3288	83.24%	1846	83.62%
4	Khmer OS Content		644	96.72%	159	98.59%
5	Khmer OS Fasthand		4605	76.53%	2012	82.15%
6	Khmer OS Freehand		2184	88.87%	505	95.52%
7	Khmer OS Metal Chrieng		3230	83.54%	1694	84.97%
8	Khmer OS Muol		525	97.32%	135	98.80%
9	Khmer OS Muol Light		3430	82.52%	2115	81.24%
10	Khmer OS Muol Pali		6697	65.86%	3688	67.28%
11	Khmer OS Siemreap		1318	93.28%	582	94.84%
12	Khmer OS System		2338	88.08%	1076	90.45%
13	Noto Serif Khmer Bold		2006	89.77%	1038	90.79%
14	Noto Serif Khmer Regular		2043	89.59%	1045	90.73%
Average :			2,454.64	87.49%	1,191.79	89.43%

No.	Font Name	Size	Character Accuracy		Cluster Accuracy	
			Errors	Accuracy	Errors	Accuracy
1	Limon F1	22	11459	41.59%	7340	34.88%
2	Limon F2		11152	43.15%	7652	32.11%
3	Limon F3		8454	56.91%	5590	50.41%
4	Limon F4		17082	12.93%	10319	8.45%
5	Limon F5		12259	37.51%	7689	31.79%
6	Limon F6		6146	68.67%	3858	65.77%
7	Limon F7		6750	65.59%	4031	64.24%
8	Limon F8		5512	71.90%	3482	69.11%
9	Limon R1		7053	64.05%	4336	61.53%
10	Limon R2		8044	59.00%	4584	59.33%
11	Limon R3		10467	46.65%	5357	52.48%
12	Limon R4		11815	39.77%	7068	37.30%
13	Limon R5		9479	51.68%	5811	48.45%
14	Limon S1		2704	86.22%	1479	86.88%
15	Limon S2		2277	88.39%	1248	88.93%
16	Limon S3		2863	85.41%	1413	87.46%
17	Limon S4		4631	76.39%	3074	72.73%
18	Limon S5		2552	86.99%	1488	86.80%
19	Limon S6		4577	76.67%	2700	76.05%
20	Limon S7		2773	85.87%	1485	86.83%
Average :			7,402.45	62.27%	4,500.20	60.08%

Accuracy Measurement and Comparison (no noise)

Engine: Fine Tune For Limon S1 Unicode

Net_spec value "[1,48,0,1 Ct3,3,16 Mp3,3 Lfys64 Lfx96 Lrx96 Lfx384 O1c1]"

Testing Text:

Characters :

19618

Clusters :

11272

No.	Font Name	Size	Character Accuracy		Cluster Accuracy	
			Errors	Accuracy	Errors	Accuracy
1	Khmer OS	12	1785	90.90%	857	92.40%
2	Khmer OS Battambang		851	95.66%	343	96.96%
3	Khmer OS Bokor		2683	86.32%	1563	86.13%
4	Khmer OS Content		879	95.52%	334	97.04%
5	Khmer OS Fasthand		6512	66.81%	3315	70.59%
6	Khmer OS Freehand		2269	88.43%	537	95.24%
7	Khmer OS Metal Chrieng		3697	81.16%	1952	82.68%
8	Khmer OS Muol		844	95.70%	369	96.73%
9	Khmer OS Muol Light		3511	82.10%	2300	79.60%
10	Khmer OS Muol Pali		5827	70.30%	2985	73.52%
11	Khmer OS Siemreap		1698	91.34%	845	92.50%
12	Khmer OS System		2595	86.77%	1254	88.88%
13	Noto Serif Khmer Bold		2220	88.68%	1207	89.29%
14	Noto Serif Khmer Regular		2276	88.40%	1217	89.20%
Average :			2,689.07	86.29%	1,362.71	87.91%

No.	Font Name	Size	Character Accuracy		Cluster Accuracy	
			Errors	Accuracy	Errors	Accuracy
1	Limon F1	22	8196	58.22%	5574	50.55%
2	Limon F2		8104	58.69%	5852	48.08%
3	Limon F3		6240	68.19%	4227	62.50%
4	Limon F4		15665	20.15%	9756	13.45%
5	Limon F5		8551	56.41%	5673	49.67%
6	Limon F6		3874	80.25%	2494	77.87%
7	Limon F7		5185	73.57%	2927	74.03%
8	Limon F8		3807	80.59%	2393	78.77%
9	Limon R1		5341	72.78%	3347	70.31%
10	Limon R2		6558	66.57%	3621	67.88%
11	Limon R3		10026	48.89%	4610	59.10%
12	Limon R4		9767	50.21%	4862	56.87%
13	Limon R5		6609	66.31%	3991	64.59%
14	Limon S1		1314	93.30%	632	94.39%
15	Limon S2		1787	90.89%	947	91.60%
16	Limon S3		2697	86.25%	1349	88.03%
17	Limon S4		2811	85.67%	1779	84.22%
18	Limon S5		3106	84.17%	1769	84.31%
19	Limon S6		3254	83.41%	1730	84.65%
20	Limon S7		2220	88.68%	1110	90.15%
Average :			5,755.60	70.66%	3,432.15	69.55%

Engine: Fine Tune For Limon S1 & F1 Unicode

Net_spec value "[1,48,0,1 Ct3,3,16 Mp3,3 Lfys64 Lfx96 Lrx96 Lfx384 O1c1]"

Testing Text:

Characters :

19618

Clusters :

11272

No.	Font Name	Size	Character Accuracy		Cluster Accuracy	
			Errors	Accuracy	Errors	Accuracy
1	Khmer OS	12	1906	90.28%	972	91.38%
2	Khmer OS Battambang		769	96.08%	276	97.55%
3	Khmer OS Bokor		3054	84.43%	1826	83.80%
4	Khmer OS Content		878	95.52%	334	97.04%
5	Khmer OS Fasthand		5845	70.21%	2857	74.65%
6	Khmer OS Freehand		2348	88.03%	637	94.35%
7	Khmer OS Metal Chrieng		4664	76.23%	2756	75.55%
8	Khmer OS Muol		1097	94.41%	575	94.90%
9	Khmer OS Muol Light		3700	81.14%	2269	79.87%
10	Khmer OS Muol Pali		6733	65.68%	3728	66.93%
11	Khmer OS Siemreap		1982	89.90%	1024	90.92%
12	Khmer OS System		2844	85.50%	1463	87.02%
13	Noto Serif Khmer Bold		2271	88.42%	1261	88.81%
14	Noto Serif Khmer Regular		2305	88.25%	1260	88.82%
Average :			2,885.43	85.29%	1,517.00	86.54%

No.	Font Name	Size	Character Accuracy		Cluster Accuracy	
			Errors	Accuracy	Errors	Accuracy
1	Limon F1	22	2209	88.74%	1400	87.58%
2	Limon F2		8981	54.22%	6293	44.17%
3	Limon F3		7172	63.44%	4785	57.55%
4	Limon F4		14651	25.32%	9305	17.45%
5	Limon F5		5306	72.95%	3531	68.67%
6	Limon F6		2536	87.07%	1577	86.01%
7	Limon F7		4109	79.05%	2119	81.20%
8	Limon F8		5063	74.19%	3226	71.38%
9	Limon R1		6461	67.07%	4092	63.70%
10	Limon R2		7637	61.07%	4541	59.71%
11	Limon R3		10578	46.08%	4893	56.59%
12	Limon R4		11410	41.84%	6259	44.47%
13	Limon R5		7859	59.94%	4851	56.96%
14	Limon S1		1560	92.05%	782	93.06%
15	Limon S2		1747	91.09%	908	91.94%
16	Limon S3		2731	86.08%	1334	88.17%
17	Limon S4		2755	85.96%	1785	84.16%
18	Limon S5		2491	87.30%	1391	87.66%
19	Limon S6		4034	79.44%	2334	79.29%
20	Limon S7		2207	88.75%	1104	90.21%
Average :			5,574.85	71.58%	3,325.50	70.50%

Accuracy Measurement and Comparison (no noise)

Engine: Fine Tune For Limon S1, S2 & F1 Unicode

Net_spec value "[1,48,0,1 Ct3,3,16 Mp3,3 Lfys64 Lfx96 Lrx96 Lfx384 O1c1]"

Testing Text:

Characters :

19618

Clusters :

11272

No.	Font Name	Size	Character Accuracy		Cluster Accuracy	
			Errors	Accuracy	Errors	Accuracy
1	Khmer OS	12	1911	90.26%	967	91.42%
2	Khmer OS Battambang		776	96.04%	291	97.42%
3	Khmer OS Bokor		2986	84.78%	1764	84.35%
4	Khmer OS Content		901	95.41%	360	96.81%
5	Khmer OS Fasthand		5617	71.37%	2664	76.37%
6	Khmer OS Freehand		2730	86.08%	841	92.54%
7	Khmer OS Metal Chrieng		3881	80.22%	2137	81.04%
8	Khmer OS Muol		1191	93.93%	657	94.17%
9	Khmer OS Muol Light		3806	80.60%	2261	79.94%
10	Khmer OS Muol Pali		6536	66.68%	3589	68.16%
11	Khmer OS Siemreap		2024	89.68%	1058	90.61%
12	Khmer OS System		2577	86.86%	1290	88.56%
13	Noto Serif Khmer Bold		2293	88.31%	1278	88.66%
14	Noto Serif Khmer Regular		2307	88.24%	1253	88.88%
Average :			2,824.00	85.60%	1,457.86	87.07%

No.	Font Name	Size	Character Accuracy		Cluster Accuracy	
			Errors	Accuracy	Errors	Accuracy
1	Limon F1	22	2263	88.46%	1432	87.30%
2	Limon F2		8842	54.93%	6217	44.85%
3	Limon F3		6411	67.32%	4308	61.78%
4	Limon F4		14365	26.78%	9144	18.88%
5	Limon F5		4717	75.96%	3222	71.42%
6	Limon F6		3269	83.34%	2012	82.15%
7	Limon F7		3946	79.89%	2045	81.86%
8	Limon F8		4442	77.36%	2772	75.41%
9	Limon R1		6001	69.41%	3829	66.03%
10	Limon R2		8083	58.80%	4761	57.76%
11	Limon R3		11133	43.25%	5080	54.93%
12	Limon R4		11854	39.58%	6614	41.32%
13	Limon R5		7378	62.39%	4565	59.50%
14	Limon S1		1252	93.62%	600	94.68%
15	Limon S2		1408	92.82%	640	94.32%
16	Limon S3		2458	87.47%	1226	89.12%
17	Limon S4		2353	88.01%	1553	86.22%
18	Limon S5		2125	89.17%	1182	89.51%
19	Limon S6		3864	80.30%	2201	80.47%
20	Limon S7		2339	88.08%	1184	89.50%
Average :			5,425.15	72.35%	3,229.35	71.35%

Engine: Fine Tune Limon S1 & S2 Unicode

(new net_spec value "[1,36,0,1 Ct3,3,16 Mp3,3 Lfys48 Lfx96 Lrx96 Lfx256 O1c111]")

Testing Text:

Characters :

19618

Clusters :

11272

Font Name	Size	Character Accuracy		Cluster Accuracy	
		Errors	Accuracy	Errors	Accuracy
Average [Khmer Unicode fonts] :	12	2,663.79	86.42%	1,333.64	88.17%

Font Name	Size	Character Accuracy		Cluster Accuracy	
		Errors	Accuracy	Errors	Accuracy
Average [Limon fonts] :	22	5,901.50	69.92%	3,524.45	68.73%

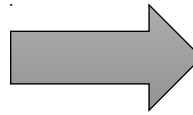
Engine: Fine Tune Limon S1 & S2 Unicode

(same net_space with Tesseract 4.0 "[1,48,0,1 Ct3,3,16 Mp3,3 Lfys64 Lfx96 Lrx96 Lfx384 O1c1]")

Font Name	Size	Character Accuracy		Cluster Accuracy	
		Errors	Accuracy	Errors	Accuracy
Average [Khmer Unicode fonts] :	12	2,815.07	85.65%	1,418.43	87.42%

Font Name	Size	Character Accuracy		Cluster Accuracy	
		Errors	Accuracy	Errors	Accuracy
Average [Limon fonts] :	22	6,154.30	68.46%	3,656.70	67.56%

ស៊ីវិលកម្ពុជា ជាភាសាធារាវងហៅថា
ge សរសេរកាត់ « A.A.C.C
ាត នៃអាកាសធរស៊ីវិល ដែលមាន
កាសធរដោយផ្អែកលើមូលដ្ឋាន នៃក



ស៊ីវិលកម្ពុជា ជាភាសាធារាវងហៅថា
09 សរសេរកាត់ « *.១.៩
ាត នៃអាកាសធរស៊ីវិល ដែលមាន
ាសធរដោយផ្អែកលើមូលដ្ឋាន នៃក
09

Testing Text:

Characters: 107

Clusters : 54

No.	Engine	Character Accuracy		Cluster Accuracy	
		Errors	Accuracy	Errors	Accuracy
1	Tesseract 4.0	29	72.90%	13	75.93%
2	Fine tune for LimonS1Unicode	15	85.98%	9	83.33%
3	Fine tune for LimonS1S2Unicode(new Nets)	19	82.24%	11	79.63%
4	Fine tune for LimonS1S2Unicode(Tesseract)	17	84.11%	10	81.48%
5	Fine tune for LimonS1F1Unicode	13	87.85%	5	90.74%
6	Fine tune for LimonS1S2F1Unicode	15	85.98%	5	90.74%

ឆ្នាំទី ១ លេខ ៩ ថ្ងៃទី ៣១ ខែ វិច្ឆិកា ឆ្នាំ ១៩៩៣

រាជកិច្ច

ចំណុះ

ភ្នំពេញ ថ្ងៃទី ១៥ ខែ វិច្ឆិកា ឆ្នាំ ៩៣

អនុក្រឹត្យលេខ ១១ អនក្រ

ស្តីពីការរៀបចំ និងការប្រព្រឹត្តទៅនៃអាជ្ញាធរ

អាកាសចរ ស្មើចលកម្ពុជា

រាជរដ្ឋាភិបាល

- បានឃើញរដ្ឋធម្មនុញ្ញ នៃព្រះរាជាណាចក្រកម្ពុជា
- បានឃើញព្រះរាជក្រឹត្យ របស់ព្រះករុណាព្រះបាទសម្តេចព្រះនរោត្តម សីហនុរ្យ័ន ព្រះចៅកម្ពុជា ស្តីពីការតែងតាំងរាជរដ្ឋាភិបាលកម្ពុជា ចុះថ្ងៃទី ០១ ខែ វិច្ឆិកា ឆ្នាំ ១៩៩៣
- យោងតាមការចាំបាច់

សម្រេច

មាត្រា ១ : អាជ្ញាធរអាកាសចរស្មើចលកម្ពុជា ជាភាសាបារាំងហៅថា **Autorité de l'Aviation Civile du Cambodge** សរសេរកាត់ “ **A.A.C.C** ” បានបង្កើតឡើងក្នុងគោលបំណងរៀបចំវិញ នូវគ្រប់សកម្មភាព នៃអាកាសចរស្មើចល ដែលមានកន្លងមកនិងឈានទៅអនុវត្តនយោបាយ សេរីរូបនីយ កម្មផ្នែកអាកាសចរដោយផ្អែកលើមូលដ្ឋាន នៃការប្រកួតប្រជែង និងមានប្រសិទ្ធភាព ។

នយោបាយសេរីរូបនីយកម្មនេះ គឺជាមូលដ្ឋានសំរាប់អនុញ្ញាតអោយផ្នែកអាកាសចរស្មើចល បើកទ្វារចំហចំពោះសកលលោកទាំងមូល ព្រមទាំងឆ្លើយតបបានទៅនឹងកំណើននៃចរាចរផ្លូវអាកាស និងចំនួនអ្នកធ្វើដំណើរត្រូវបានឡើង ។

ការដាក់ផ្នែកនេះអោយមានសេវាកម្មដ៏ត្រឹមត្រូវលើផ្នែកបច្ចេកទេស ប្រកបដោយប្រសិទ្ធភាព គឺចាំបាច់ដាច់ខាត សំដៅទៅទាញក្រុមហ៊ុន អាកាសចរជាច្រើនអោយចូលមកកម្ពុជា ។

មាត្រា ២ : អាជ្ញាធរអាកាសចរស្មើចលកម្ពុជា មានបេសកកម្មដូចខាងក្រោម ៖

- អាជ្ញាធរអាកាសចរស្មើចលកម្ពុជា ស្ថិតនៅក្រោមការគ្រប់គ្រងផ្ទាល់របស់ទីស្តីការគណៈរដ្ឋមន្ត្រី ។
- អាជ្ញាធរអាកាសចរស្មើចលកម្ពុជា ដឹកនាំដោយប្រធានមួយរូបមានឋានៈជាអនុរដ្ឋលេខាធិការសមាជិកមួយចំនួនទៀត ដែលត្រូវតែងតាំងឡើងដោយរាជរដ្ឋាភិបាលកម្ពុជា ។

ជោទី ១ លេខ ៩ ថ្ងៃទី ១ ខែ វិច្ឆិកា ឆ្នាំ ១៩៩៣ រាជកិច្ច ទំព័រ

ល ខ = ៦ -

ក្នុងពេល ថ្ងៃទី ១៩ 48 រំលឹក នាំ ០ :

រងគកើតលេ អខ8រត

អនុក្រឹត្យលេខ ១១ អនក្រ

ជល ដ ង 2 22 ។ គ

ស្តុពលការរៀបចំ និងការប្រព្រឹត្តទៅនេះអាជ្ញាធរ

+ ៣ »

អាកាសចរ ស្មើចលកម្ពុជា

4

រាជរដ្ឋាភិបាល

- បានឃើញរដ្ឋធម្មនុញ្ញ នៃព្រះរាជាណាចក្រកម្ពុជា

បានឃើញព្រះរាជក្រឹត្យ របស់ព្រះករុណាព្រះបាទសម្តេចព្រះនរោត្តម សីហនុរ្យ័ន ព្រះចៅកម្ពុជាស្តីពីការតែងតាំងរាជរដ្ឋាភិបាលកម្ពុជា ចុះថ្ងៃទី ០១ ខែ វិច្ឆិកា ឆ្នាំ១៩៩៣

យោងតាមការចាំបាច់

សម្រេច

មាត្រា ២ អាជ្ញាធរអាកាសចរស្មើចលកម្ពុជា ជាភាសាបារាំងហៅថា ២០1148 ៥8 ក្រុមរៀបចំ ៨០ ជប ០២កង០៨ សរសេរកាត់ ។ .គ.០.០ * បានបង្កើតឡើងក្នុងគោលបំណងរៀបចំវិញ នូវគ្រប់សកម្មភាព នៃអាកាសចរស្មើចល ដែលមានកន្លងមកនិងឈានទៅអនុវត្តនយោបាយ សេរីរូបនីយ កម្មផ្នែកអាកាសចរដោយផ្អែកលើមូលដ្ឋាន នៃការប្រកួតប្រជែង និងមានប្រសិទ្ធភាព ។

នយោបាយសេរីរូបនីយកម្មនេះ ន គឺជាមូលដ្ឋានសំរាប់អនុញ្ញាតអោយផ្នែកអាកាសចរស្មើចល/ បើកទ្វារចំហចំពោះសកលលោកទាំងមូល ព្រមទាំងឆ្លើយតបបានទៅនឹងកំណើននៃចរាចរផ្លូវអាកាស

Image 2

Testing Text:

Characters: 1315

Clusters : 737

No.	Engine	Character Accuracy		Cluster Accuracy	
		Errors	Accuracy	Errors	Accuracy
1	Tesseract 4.0	207	84.26%	109	85.21%
2	Fine tune for LimonS1Unicode	260	80.23%	132	82.09%
3	Fine tune for LimonS1S2Unicode(new Nets)	203	84.56%	82	88.87%
4	Fine tune for LimonS1S2Unicode(Tesseract)	209	84.11%	90	87.79%
5	Fine tune for LimonS1F1Unicode	227	82.74%	112	84.80%
6	Fine tune for LimonS1S2F1Unicode	240	81.75%	99	86.57%

ទី ១៤ លេខ ០១

សប្តាហ៍ទី ១ ថ្ងៃទី ០៤ ខែមករា ឆ្នាំ ២០១៤

រាជកិច្ច ទំព័រ ១១

3+

៣



ព្រះរាជក្រឹត្យ
យើង

នស / រកត / ០២០១ / ១៣៤៦

១៤ លេខ ០១ សប្តាហ៍ទី ១ ថ្ងៃទី ០៤ ខែមករា ឆ្នាំ ២០១៤ រាជកិច្ច ទំព័រ ១១

ព្រះករុណាព្រះបាទសម្តេចព្រះបរមនាថ នរោត្តម សីហមុនី
សមានភូមិជាតិសាសនា រក្ខតខត្តិយា រមរាជវង្សា ពូជីន្រ្ទាទាមហាគរ្យត្រ
ខេមរាជនា សមូហោភាស គម្ពុជឯករាជរដ្ឋបូរណសន្តិ សូតមង្គលា សិរីចិប្បលា
ខេមរាស្រីពិរាស្ត្រ ព្រះចៅក្រុងកម្ពុជាធិបតី

ព្រះរាជក្រឹត្យ

យើង

ព្រះករុណាព្រះបាទសម្តេចព្រះបរមនាថ នរោត្តម សីហមុនី

សមានភូមិជាតិសាសនា រក្ខតខត្តិយា រមរាជវង្សា ពូជីន្រ្ទាទាមហាគរ្យត្រ

ខេមរាជនា សមូហោភាស គម្ពុជឯករាជរដ្ឋបូរណសន្តិ សូតមង្គលា សិរីចិប្បលា
ខេមរាស្រីពិរាស្ត្រ ព្រះចៅក្រុងកម្ពុជាធិបតី

នស រកត / ០២០១០៣

បានទ្រង់យល់រដ្ឋធម្មនុញ្ញ នៃព្រះរាជាណាចក្រកម្ពុជា

បានទ្រង់យល់ព្រះរាជក្រឹត្យ លេខ នស/រកត/០៩១/៩០ ចុះថ្ងៃទី ២៤ ខែ កញ្ញា ឆ្នាំ ២០

- បានទ្រង់យល់រដ្ឋធម្មនុញ្ញ នៃព្រះរាជាណាចក្រកម្ពុជា
- បានទ្រង់យល់ព្រះរាជក្រឹត្យ លេខ នស/រកត/០៩១/៩០៣ ចុះថ្ងៃទី ២៤ ខែ កញ្ញា ឆ្នាំ ២០១៣ ស្តីពីការតែងតាំងរាជរដ្ឋាភិបាលនៃព្រះរាជាណាចក្រកម្ពុជា
- បានទ្រង់យល់ព្រះរាជក្រម លេខ ០២/នស/៩៤ ចុះថ្ងៃទី ២០ ខែ កក្កដា ឆ្នាំ ១៩៩៤ ដែលប្រកាសឱ្យប្រើច្បាប់ ស្តីពីការរៀបចំ និងការប្រព្រឹត្តទៅនៃគណៈរដ្ឋមន្ត្រី
- បានទ្រង់យល់ព្រះរាជក្រម លេខ នស/រកម/០១៩៦/០៩ ចុះថ្ងៃទី ២៤ ខែ មករា ឆ្នាំ ១៩៩៦ ដែលប្រកាសឱ្យប្រើច្បាប់ ស្តីពីការបង្កើតទីស្តីការគណៈរដ្ឋមន្ត្រី
- បានទ្រង់យល់ព្រះរាជក្រឹត្យ លេខ នស/រកត/០៨០១/២៦៤ ចុះថ្ងៃទី ០៧ ខែ សីហា ឆ្នាំ ២០០១ ស្តីពីនីតិសារជាតិនៃមជ្ឈមណ្ឌលសកម្មភាពកំចាត់មីនកម្ពុជា
- បានទ្រង់យល់ព្រះរាជក្រឹត្យ លេខ នស/រកត/០១០៨/០៧៨ ចុះថ្ងៃទី ៣១ ខែ មករា ឆ្នាំ ២០០៨ ស្តីពីវិសោធនកម្មមាត្រា ៨, មាត្រា ៩ និងមាត្រា ១១ ជំពូកទី ៤ នៃព្រះរាជក្រឹត្យ លេខ នស/រកត/០៨០១/២៦៤ ចុះថ្ងៃទី ០៧ ខែ សីហា ឆ្នាំ ២០០១ ស្តីពីនីតិសារជាតិនៃមជ្ឈមណ្ឌលសកម្មភាពកំចាត់មីនកម្ពុជា
- បានទ្រង់យល់សេចក្តីក្រាបបង្គំទូលថ្វាយរបស់សម្តេចអគ្គមហាសេនាបតីតេជោ **ហ៊ុន សែន** នាយករដ្ឋមន្ត្រី នៃព្រះរាជាណាចក្រកម្ពុជា

ត្រាស់បង្គាប់

មាត្រា ១: តែងតាំងសមាសភាពប្រធាន និងអនុប្រធានក្រុមប្រឹក្សាភិបាល នៃមជ្ឈមណ្ឌលសកម្មភាពកំចាត់មីនកម្ពុជា ៖

- ១- ឯកឧត្តម **វិម សោភ័ណ** ជាប្រធានរដ្ឋសិទ្ធិក្រុមប្រឹក្សាភិបាល មានឋានៈស្មើ ទេសរដ្ឋមន្ត្រី
- ២- ឯកឧត្តម **ឡេង សុផា** ជាអនុប្រធានអចិន្ត្រៃយ៍ក្រុមប្រឹក្សាភិបាល មានឋានៈស្មើ រដ្ឋមន្ត្រី ។

មាត្រា ២: សម្តេចអគ្គមហាសេនាបតីតេជោ **ហ៊ុន សែន** នាយករដ្ឋមន្ត្រី នៃព្រះរាជាណាចក្រកម្ពុជា ត្រូវទទួលបន្ទុកអនុវត្តព្រះរាជក្រឹត្យនេះ ចាប់ពីថ្ងៃចេញព្រះហស្តលេខតទៅ ។

ព្រះហស្តលេខា និងព្រះរាជលញ្ឆករ
នរោត្តម សីហមុនី

ពល. ១៣១២.១៣៤៦

លេខ ១៣១៣ ស.ណ
ចំណុះនៃព្រះរាជាណាចក្រកម្ពុជា

Image 3

Testing Text:		Characters: 1951		Clusters : 1025	
No.	Engine	Character Accuracy		Cluster Accuracy	
		Errors	Accuracy	Errors	Accuracy
1	Tesseract 4.0	267	86.31%	139	86.44%
2	Fine tune for LimonS1Unicode	256	86.88%	112	89.07%
3	Fine tune for LimonS1S2Unicode(new Nets)	300	84.62%	142	86.15%
4	Fine tune for LimonS1S2Unicode(Tesseract)	296	84.83%	147	85.66%
5	Fine tune for LimonS1F1Unicode	279	85.70%	140	86.34%
6	Fine tune for LimonS1S2F1Unicode	271	86.11%	127	87.61%

ការស្វែងយល់អំពីបញ្ហាទុរភាពភាពនៃជនជាតិដើម

និងការមូលហេតុនៃការយល់ដឹង

Understanding the Poverty Concerns of Indigenous Peoples:
Trends, Causes and Perceptions

បណ្ឌិត ហ៊ាន សុទុច

វិទ្យាស្ថានសិក្សាស្រាវជ្រាវថាវ៉ាន់ខ្ពស់

ជំពូក១: ជនជាតិដើមភាគតិចនៅកម្ពុជា

១. ប្រជាជនគោលដៅ

ទូរស័ព្ទនិងទូរសារ: ០២៣ ២១៤

សារអេឡិចត្រូនិច: sokhom@forum.org.kh or cas@forum.org.kh

ការស្វែងយល់អំពីបញ្ហាទុរភាពភាពនៃជនជាតិដើម

និងការមូលហេតុនៃការយល់ដឹង

៤៤៨ខក ជា8 7018ក៤៧ 000៧ ទៅ ៤៨!១0២0៤ ៤១0៦83

វិទ្យាស្ថានសិក្សាស្រាវជ្រាវថាវ៉ាន់ខ្ពស់ 83៤383 2៧៨ ៦88៦៥003

ថាទ សុទុច / ៩ ទូរស័ព្ទនិងទូរសារ ០: ០២ ២១៤ ៤៩៤

វិទ្យាស្ថានសិក្សាស្រាវជ្រាវថាវ៉ាន់ខ្ពស់ សារអេឡិចត្រូនិច ០: 304/277/៦/សនា.0.4៣ 07 488ស៦៥សក.02១២.ស

ជំពូក១ ជនជាតិដើមភាគតិចនៅកម្ពុជា

១. ប្រជាជនគោលដៅ

គេប្រើពាក្យផ្សេងៗគ្នាក្នុងការហៅនិងពិណ្ឌនាជនជាតិដើមភាគតិចដែលរស់នៅតំបន់ឥសាន និងតំបន់ឧត្តរភាគនៃប្រទេសកម្ពុជា ដូចជា៖ "ជនជាតិភាគតិច" "កុលសម្ព័ន្ធអ្នា" "អ្នករស់តាមតំបន់ខ្ពង់រាប" "ជនជាតិក្រៅស្រុក" "ជនជាតិដើម"។ ហើយប្រជាពលរដ្ឋទាំងនោះតែងតែហៅខ្លួនឯងថា "ជនជាតិ"។ ដូចនេះដើម្បីងាយស្រួលយើងនឹងប្រើពាក្យថា "ជនជាតិដើម" ឬ "ជនជាតិតាមតំបន់ខ្ពង់រាប" ជាជាងប្រើពាក្យ "ជនជាតិភាគតិច" ឬ "ជនជាតិភាគតិច"។ ប្រជាពលរដ្ឋទាំងនោះភាគច្រើនរស់នៅក្នុងតំបន់ខ្ពង់រាបភាគឥសាន គឺខេត្តរតនគិរី មណ្ឌលគីរី ក្រចេះនិងស្ទឹងត្រែង។

Image 4

Testing Text:

Characters: 807

Clusters : 389

No.	Engine	Character Accuracy		Cluster Accuracy	
		Errors	Accuracy	Errors	Accuracy
1	Tesseract 4.0	225	72.12%	64	83.55%
2	Fine tune for LimonS1Unicode	268	66.79%	87	77.63%
3	Fine tune for LimonS1S2Unicode(new Nets)	245	69.64%	76	80.46%
4	Fine tune for LimonS1S2Unicode(Tesseract)	233	71.13%	59	84.83%
5	Fine tune for LimonS1F1Unicode	212	73.73%	52	86.63%
6	Fine tune for LimonS1S2F1Unicode	228	71.75%	54	86.12%

Conclusion:

The Accuracy of Legacy Limon fonts could be improved by fine tuning the engine for more Limon fonts.

Future Work:

- Fine tune for all Limon Legacy fonts**
- Experiment with different net_spec and Tesseract parameters**

The background of the slide is a photograph of a winding asphalt road at sunset or sunrise. The sky and landscape are bathed in a warm, orange-gold light. A small car is visible in the distance on the road. Overlaid on the image are several semi-transparent geometric shapes: a purple triangle on the left, a large orange shape on the right, and a light blue and green shape at the bottom right.

**Thanks for your
attention !!**