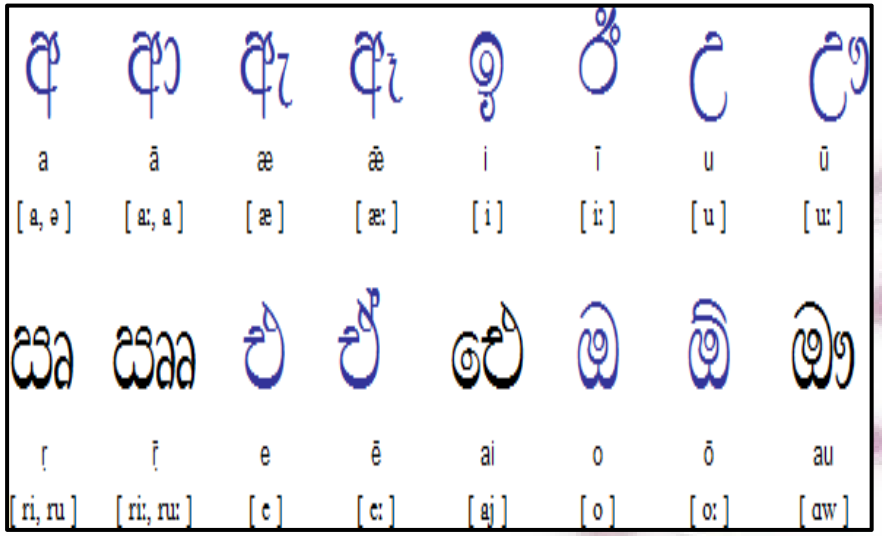
OPTICAL CHARACTER RECOGNITION

SINHALA LANGUAGE

Sinhala Script consists of:

18 vowels



40 consonants

ක	ka	ඛ	kha	ග	ga	ඝ	gha	ඞ	nga	ඟ	nnga				
ච	ca	ඡ	cha	ජ	ja	ඣ	jha	ඤ	nya	ඦ	jnya	ඥ	nyja		
ට	tta	ඨ	ttha	ඩ	dda	ඪ	ddha	ණ	nna	ඬ	nndda				
ත	ta	ථ	tha	ද	da	ධ	dha	න	na	ඳ	nda				
ප	pa	ඵ	pha	බ	ba	භ	bha	ම	ma	ඹ	mba				
ය	ya	ර	ra	ල	la	ව	va								
ශ	sha	ෂ	ssa	ස	sa	හ	ha	ළ	lla	ෆ	fa				

Sinhala Script

18 modifiers

other symbols (rakaranshaya, yansaya)

Font: FMAbhaya

Font Size :12

Recognition results - Decision Tree

From the connected components, training and test data sets were created.

Overall accuracy:

92.61%

Type	Total_Items	Correctly_Identified	Accuracy
500	15	15	100
501	15	15	100
502	15	15	100
503	15	15	100
504	15	15	100
505	15	15	100
506	15	9	60
507	15	14	93
508	15	15	100
509	15	15	100
510	15	15	100
511	15	13	86
512	16	13	81
513	15	15	100
514	15	15	100
515	15	10	66
516	15	14	93
517	15	14	93

Recognition results - Tesseract

Overall accuracy:

95.20%

Charcode	test samples	Correctlyidentified	Accuracy %
A00500	15	14	93.33
A00501	15	15	100.00
A00502	15	15	100.00
A00503	15	15	100.00
A00504	15	15	100.00
A00505	15	14	93.33
A00506	15	15	100.00
A00507	15	15	100.00
A00508	15	15	100.00
A00509	15	15	100.00
A00510	15	4	26.67
A00511	15	15	100.00
A00512	16	16	100.00
A00513	15	15	100.00
A00514	15	15	100.00
A00515	15	15	100.00
A00516	15	15	100.00
A00517	15	15	100.00

Image Document

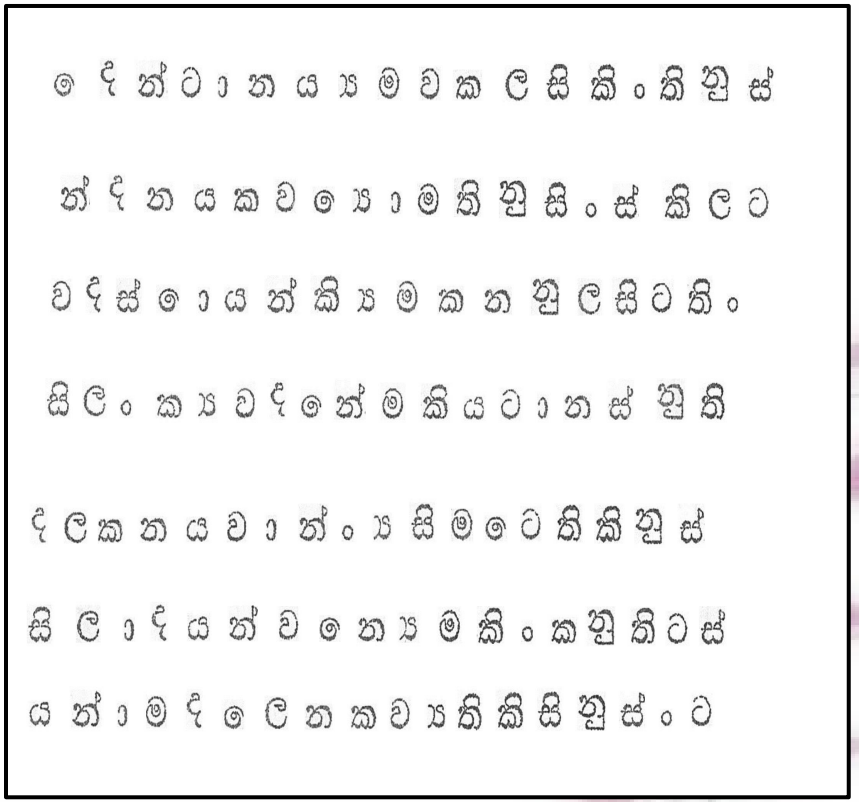


Image document has 18 different character types and 7 samples of each character type.

Full OCR - Decision Tree method

sample OUTPUT of line segmentation

all the lines were segmented properly.

Accuracy 100%

Output Full OCR of DT Method

Line	Total Chars	Correctly identified	accuracy
1	18	3	16.67
2	18	5	27.78
3	18	5	27.78
4	18	6	33.33
5	18	4	22.22
6	18	5	27.78
7	18	4	22.22
overall	126	32	25.40

Manually counted

Overall Accuracy

25.4%

Output of Full OCR (mixed chars) - Tesseract

outputsin.txt

```
1  A00506A00502A00503A00513A00500A00509A00516A00517A00508A00515A00504A00507A00512A00505A00501A00514A00510A00511
2  A00503A00502A00509A00516A00504A00515A00506A00517A00501A00508A00514A00510A00512A00501A00511A00505A00507A00513
3  A00515A00502A00511A00506A00500A00516A00503A00505A00517A00508A00504A00509A00510A00507A00512A00513A00514A00501
4  A00512A00507A00501A00504A00517A00515A00502A00506A00503 A00508A00505A00516A00513A00500A00509A00511A00510A00514
5  A00502A00507A00504A00509A00516A00515A00500A00503A00501A00517A00512A00508A00506A00513A00514A00505A00510A00511
6  A00512A00507A00500A00502A00516A00503A00515A00506A00509A00517A00508 A00505A00501A00504A00510A00514A00513A00511
7  A00516A00503A00500A00508A00502A00506A00507A00509A00504A00515A00517A00514A00505A00512A00510A00511A00501A00513
```

7 lines of text identified. Each line contains 18 characters.

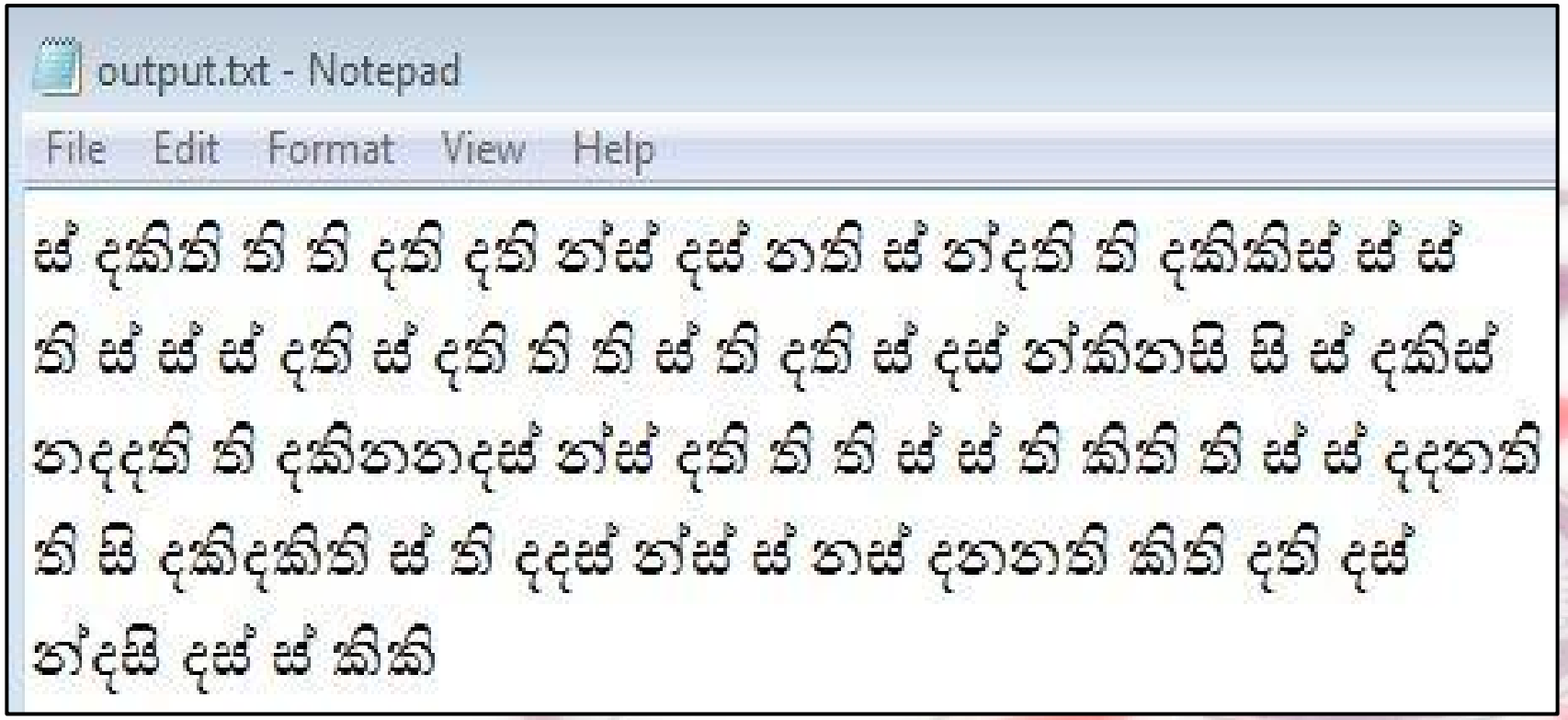
Full OCR - Results and accuracy

Line	Total Chars	Correctly identified	accuracy
1	18	18	100.00
2	18	17	94.44
3	18	18	100.00
4	18	18	100.00
5	18	18	100.00
6	18	18	100.00
7	18	18	100.00
overall	126	125	99.21

Manually counted accuracy values are presented in the table

Overall Accuracy 99.21%

Full OCR- Final Output by DT Algorithm



Group members

- Malinda Punchimudiyanse
- Aruna Lorensuhewa
- Dhanika Perera
- Hiroshi de Silva