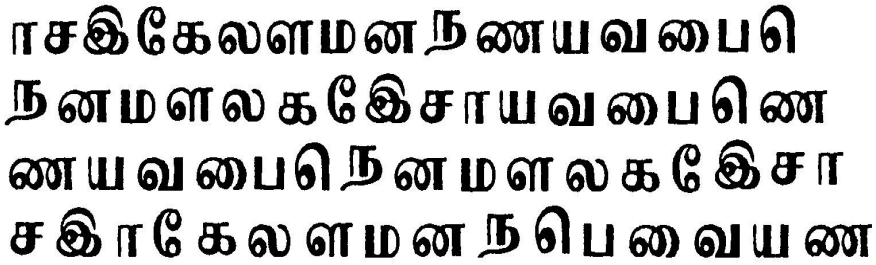
Optical Character Recognition of Tamil Script



Data set 1:

Tesseract	**:**	**90.8%**
Decision Tree	**:**	**72.5%**

Data set 2:

Decision Tree	**:**	**<20%**

K. Sarveswaran

M.H.Minver

Syllables

Syllable String	Main body ID
	500
	501
	502
	503
	504

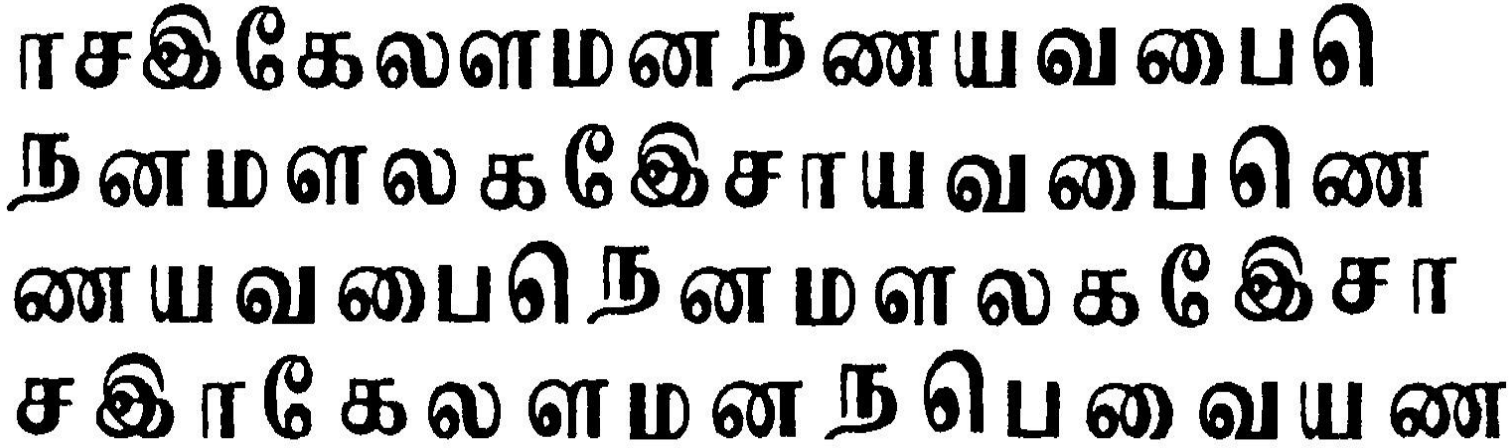
Syllable String	Main body ID
	505
	506
	507
	508
	509

Syllable String	Main body ID
	511
	512
	513
	514
	515
	516

Data set 1

Main body Type	Total tokens in document images	Total unique syllables in document images
16	240	16

Data set 2



Line Segmentation

Line:- 1

Line:- 2

Line:- 3

Line:- 4

Ligature segmentation

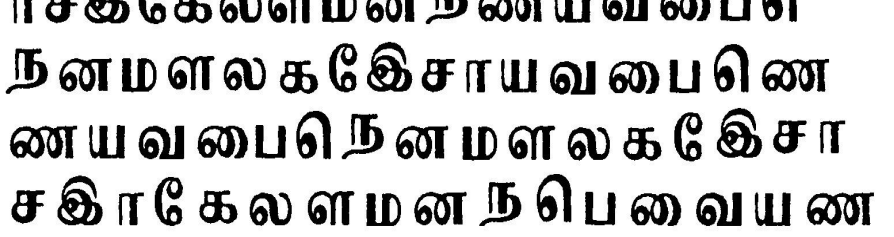
Classification and recognition results for Data set 1 using DT

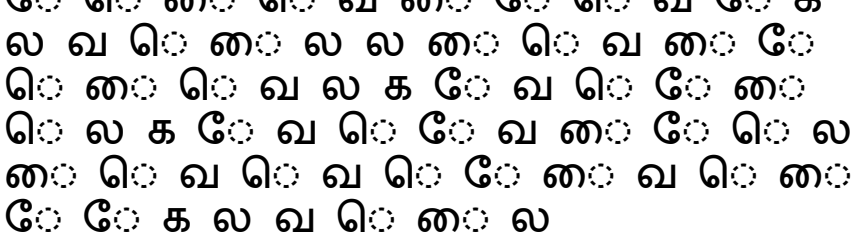
Type	T_Items	Correct	Accuracy
500	15	15	100
501	15	10	66
502	15	14	93
503	15	15	100
504	15	15	100
505	15	15	100
506	15	0	0
507	15	15	100
508	15	15	100
509	15	0	0
511	15	15	100
512	15	15	100
513	15	15	100
514	15	15	100
515	15	0	0
516	15	0	0
Total	240	174	72.5

Classification and recognition results for Data set 1 using Tesseract

Type	T_Items	Correct	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	15	100
507	15	15	100
508	15	15	100
509	15	8	72.2
511	15	15	100
512	15	15	100
513	15	15	100
514	15	15	100
515	15	15	100
516	15	0	0
Total	240	218	90.8

Data set 2 using line segmentation, ligature segmentation and DT





آپ کا شکریہ

धन्यवाद

ಧನ್ಯವಾದ

நன்றி