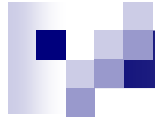




Sinhala OCR

Dulip Herath

Language Technology Research Laboratory
University of Colombo School of Computing
Sri Lanka



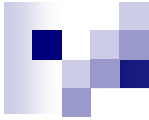
Overview

- Sinhala Script
- OCR Process
- Preprocessing
- Recognition
- Postprocessing
- Results
- Future Directions

Sinhala Script

- Sinhala script has:
 - 18 vowels
 - 40 consonants
 - 18 modifiers (including *halant*)
 - ~10 ligatures
 - Other symbols (*rephaya*, *rakarasaya*, *yansaya*)
- Typical Sinhala script has 3 layers





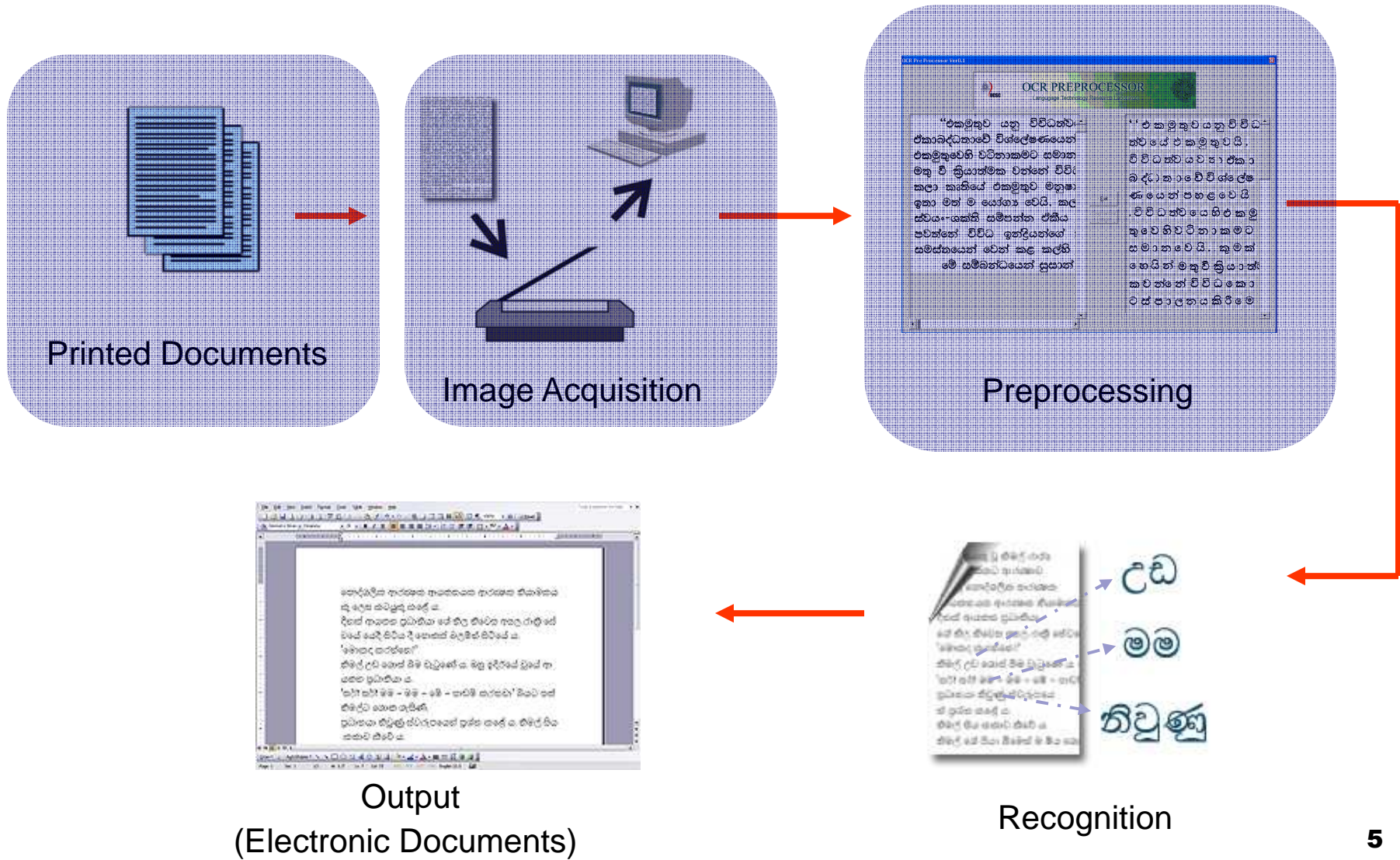
Sinhala Script

අ , ආ, ඇ, ඈ, ඉ, ඊ, උ, ඌ, ඍ, ඎ, ඏ, ඐ, එ, ඒ, ඔ, ඔ්, ඖ, ඗

ක, ක්, ග, ස, ඩ, හ, ව, ඡ, ඣ, ඤ, ඦ, ට, ඨ, ඩ්, ඩ්, ණ , ඬ, ථ, ධ, ඳ, ධ්,
න, ඳ, ප,ඵ ,බ් ,හ ,ම , ම්, ය, ර, ල, ව, ග, ඞ, ස, හ , ල, ෆ

ඒ, ඔ, ට්, ට්, ඒ, ඒ, ට්, ට්, ට්,ඔ,ඔ,ඔ,ඔ, ඔ, ඔ, ඔ, ඔ, ඔ, ඔ

OCR Process





Preprocessing

- Preprocessing stage has several tasks to be done:
 - Binarization (Dynamic Thresholding)
 - Skew Detection and Correction (Principle Component Analysis)
 - Noise Removal (Gaussian Filter)
 - Segmentation (Projection Profiles + Connected Components)
 - Normalization (25X25)

Segmentation

- Cases cannot be handled with Projection Profiles were handled by using Connected Components
 - If a segment width is greater than the average character width then apply a coloring algorithm to color connected components of the segment with each connected component with a different color
 - If the number of connected components is greater than 1 then apply vertical project for each color to determine the component boundaries





Recognition

- Recognition is done by using k-nearest neighbor (k-NN) algorithm.
 - k-NN is a memory based method
 - Maintain a set of pre-computed templates
 - Defined a similarity measure (Euclidean Distance) to find the closest template category



Postprocessing

- Having recognized each symbol:
 - Characters were reordered in order to comply with Unicode representations
 - Use information about spaces and line to present the final output similar to the source image



Results

- Skew Detection and Correction

Actual Angle	Estimated Angle	
	Mean Value	Median Value
18	17.847	17.983
10	9.740	9.996
5	4.874	5.026
3	2.784	2.983
-3	-3.009	-3.015
-5	-5.019	-5.036
-10	-10.008	-10.008
-18	-18.279	-18.064



Results

- Recognition Results

%	<i>Abhaya</i>	<i>Manel</i>	<i>Lakbima</i>	<i>Divaina</i>	<i>Letter press</i>
Recognized	97.17	96.26	89.89	78.26	95.81

Results

■ Error Analysis

Character	Confused Character	%
ක්	ක	14.29
. (period)	- (hyphen)	61.32
ශ්‍රී	ශ්‍ර	33.33
ඊ	ඊ	66.67
ඒ	ඒ	52.63
න්	න	18.18
භ	භ	28.57
භ	භ	17.64
ච	ච	8.33
ච	ච	20.00



Future Directions

- Make the system more font independent
- Improve the noise removal component
- Try different recognition algorithms such as ANNs, HMMs
- Incorporate a character n-gram model to the postprocessing module



Thank You!!