PAN Localization Project

Project No: 102042 Ref. No: PANL10n\SL\Ocrpre\UG\001



User Guide for Preprocessing Engine of the Optical Character Recognition System for Sinhala Scripts version 0.1

30th April 2005

PAN Localization Project Language Technology Research Laboratory University of Colombo School of Computing Sri Lanka

Table of Contents

1.	Introduction	3
2.	Running the Application	3
3.	Using Application	5

1 Introduction

This document explains how to use OCR Preprocessor v0.1. It is an application that consists of basic image processing techniques such as image segmentation and noise removal. The current version of OCR Preprocessor works only for the images in bitmap format.

2 Running the Application

After the successful installation of OCR preprocessor v0.1 (Refer installation guide for OCR Preprocessor v0.1 to install the application) and all dependencies, follow the steps below to start the application.

 Run the application from Windows start menu at location Start -> All Programs -> OCR Preprocessor -> OCR Preprocessor illustrated in Figure 2.1. If the link for "OCR preprocessor v0.1" is missing in the start menu please reinstall the application.

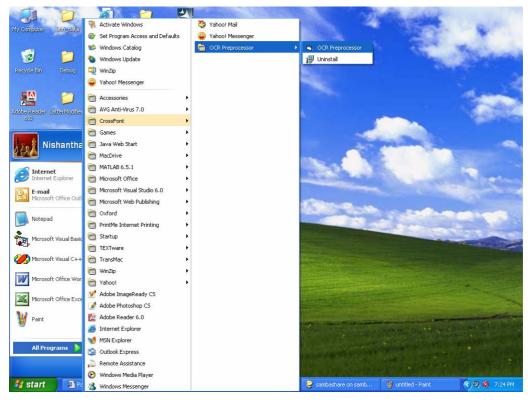


Figure 2.1: Running the "OCR Preprocessor v0.1" from Windows start menu

II. As the application starts, it shows main window as shown in Figure 2.2.

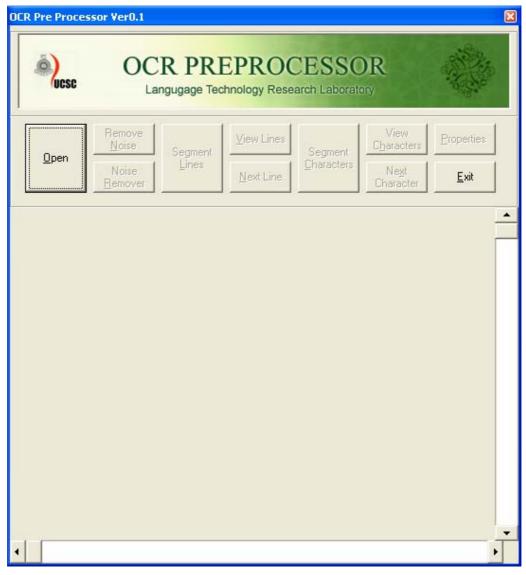


Figure 2.2: OCR Preprocessor v0.1, at start up

PAN localization project

3 Using Application

3.1 Open the source image

In order to open the image follows the steps given below.

I. Click the "Open" button. Then the open window will appear as shown in Figure 3.1.1

OCR Pre Processo	r Ver0.1			×		
OCR PREPROCESSOR Langugage Technology Research Laboratory						
<u>Open</u>	Remove <u>N</u> oise Segmen	t <u>V</u> iew Lines Segment	View Characters	ies		
Open				? 🛛		
Look jn:	🖘 Local Disk (C:)	•	+ 🗈 💣 🎟 -			
My Recent Documents Desktop My Documents	DMedagoda Documents and Se j2sdk1.4.2_06 JDK 1.4.2 MATLAB6p5p1 New Folder OCRPrepocessor Program Files Test WINDOW5 MINDOW5 Data14 FileRename	E FileRename Manel B.M copy.res Manel B.M copy.res Micdata Nicdata Node Node Node Node Process Process Process request request request	E ReSchedule rt_test1 rt_test3 Schedule			
(File <u>n</u> ame:		_	<u>O</u> pen		
My Network Places	Files of type:		•	Cancel		
FidUes	Г	Open as <u>r</u> ead-only				

Figure 3.1.1: Preprocessor v0.1, open the image

- II. Browse the image file to be opened.
- III. Then image file will appear as shown in Figure 3.1.2

OCR Pre Processor Ver0.1	×
OCR PREPROCESSOR Langugage Technology Research Laboratory	and the second sec
Image: Depen Remove Noise Segment Lines View Lines Segment Characters View Characters Image: Depen Noise Remover Next Line Next Line Next Characters Next Characters	<u>Properties</u>
"	A
''එකමුතුව යනු විවිධත්වයේ	231.
ඒකාබද්ධතාවේ විශ්ලේෂණයෙන් පහ	ළ වෙදි
එකමුතුවෙහි වටිතාකමට සමාත වෙ	යි. කම
මතු වී කිුයාත්මක වත්තේ විවිධ න	50 70 5
කලා කෘතියේ එකමුතුව මනුෂා ශ්	ර්රයෙහි
ඉතා මත් ම යෝගා වෙයි. කලා ක	ෘතියක්
ස්වය - ශක්ති සම්පන්න ඒකීය සමස	
	2022
පවත්තේ විවිධ ඉන්දියන්ගේ කිුයා	කාරත්
සමස්තයෙන් වෙන් කළ කල්හි පුන	ෂ්ට වේ
	N- +
	•

Figure 3.1.2: Image file

3.2 Removing Noise in the image

In order to remove noise in the image follows the steps given below.

I. Click the "Remove Noise " button, then filtered image appears as shown in the Figure 3.1.3

OCR Pre Proces	sor Ver0.1	×
(increased in the second secon	OCR PREPROCESSOR Langugage Technology Research Laboratory	
<u>O</u> pen	Remove Noise Segment Lines View Lines View Characters Properties Noise Remover Next Line Next Characters Next Character Exit	
		-
	එකමුතුව යනු විවිධත්වයේ එකමුතු	
ඒකාබ	ද්ධතාවේ විශ්ලේෂණයෙන් පහළ වෙරි	
	ෘවෙහි වටිතාකමට සමාත වෙයි. කුම	
මතු වී) කියාත්මක වත්තේ විවිධ කොටස්	
19.55	කෘතියේ එකමුතුව මනුෂා ශරීරයෙහි	
2010/00/00/00/00/00/00/00/00/00/00/00/00/		
ඉතා (මත් ම යෝගා වෙයි. කලා කෘතියක්	
ස්වයං	-ශක්ති සම්පන්න ඒකීය සමස්නයකි.	8
පවත්	තේ විවිධ ඉන්දියන්ගේ කියාකාරිත්	
සමස්ද	ායෙන් වෙන් කළ කල්හි පුන ෂ්ට වෙ	
		-
•		•

Figure 3.1.3: Filtered Image

II. Click "Noise Remover" to select the parameters for customize filtering. Enter the values of the parameters as shown in Figure 3.1.4.

🛢 Select Filtering Parameters 🔳 🗖 🗙						
Mask Size	7×7 💌					
Standard Devi:	ation 0.8					
ок	Cancel					

Figure 3.1.4: Parameters for filtering

III. Click "OK" to remove the noise.

3.3 Segment the image into lines

To segment the image into lines follow the steps given below.

I. Click "Segment Lines" then a dialog box appears as the Figure 3.3.1

OCR Pre Proces	sor Ver0.1	×
	OCR PREPROCESSOR Langugage Technology Research Laboratory	
<u>O</u> pen	Segment Segment Segment Segment	perties <u>E</u> xit
ඒකාබ එකමු මතු දි කලා ඉතා ස්වයං	 DMedagoda Documents and Settings j2sdk1.4.2_06 JDK 1.4.2 MATLAB6p5p1 New Folder OCRPrepocessor Program Files Test WINDDWS Solution 	ාමුඅ වෙරි කුම ටස් යක් යක් යක්.
පවත් සමස්		රිත්ෑ වෙ
•		۱.

Figure 3.3.1: Select a folder to save segmented images

II. Select a folder to save the segmented image or you can create a new folder. Click "OK" to segment the image into lines. Then it gives a massage as in the Figure 3.3.2

OCR Pre Proces	sor Ver0.1	×
	OCR PREPROCESSOR Langugage Technology Research Laboratory	
<u>O</u> pen	Remove Vew Noise Segment	<u>P</u> roperties
	Select a folder to save segmented images	<u>E</u> xit
	🖃 c: 💽	· ·
1		කමුතු
ඒකාබ	PreprocessorVer1.0	වෙරි
එකමු	The segmented files were successfuly saved in C:\Test\	. කුම
මතු ව්		ාටස්
කලා	Test	රයෙහි
ඉතා		බියක්
ස්වයං		ායකි.
පවත්	Note: Segmentation will take a while till whole file is segmented.	ාාරිත්ෑ
සමස්	Save Cancel Make New Folder	ට වේ
•		•

Figure 3.3.2: Image is segmented into lines

PAN localization project

3.4 View Lines

Click "View Lines" button then the line number 0 appears in Figure 3.4.1

OCR Pre Proces	sor Ver0.1					٤
(a)	OCR PREPROCESSOR Langugage Technology Research Laboratory					
<u>D</u> pen	Remove <u>N</u> oise <u>R</u> emover	Segment Lines	⊻iew Lines <u>N</u> ext Line	Segment <u>C</u> haracters	View Characters Negt Character	Properties Exit
"එz	කමුතුව	යනු	විවිට	ාන්වගේ	ය් එක	ාමතව
•						•

Figure 3.4.1: View Lines

3.5 Segment Lines into Characters

In order to segment the lines into characters, follow the steps given below.

- I. Click the "Segment Characters" button
- II. Select a folder to save the segmented image or create a new folder.
- III. Click "OK" button to segment the image to characters. When the segmentation is completed a massage appears as shown in Figure 3.5.1

OCR Pre Processor	Ver0.1	×
W UCSC	OCR PREPROCESSOR Langugage Technology Research Laboratory	
<u>O</u> pen	Noise Segment ⊻iew Lines Segment View Noise Lines Lines Lines Next Save the segmented characters X X X	<u>Properties</u>
"ඒ ඒකාබද්	d: [NEW VOLUME] C:\ CRPrepocessor)කමුතු: 5 වෙයි
එකමුතුං මතු වී කලා ස	done	්කුමෑ ාටස් ෑ රයෙහි
ඉතා ම ස්වය	<u>OK</u>	තියක් තයකි.
පවත්තෙ සමස්ත	Note: Segmentation will take a while till whole file is segmented. Segmented Save Cancel Make New Folder	කාරිත්ව ්ට වො ,

Figure 3.5.1: Segment Characters

IV. Click "Segmented Image" then the segmented characters appear as shown in Figure 3.5.2

OCR Pre Processor Ver0.1 (IICSC OCR PREPROCESSOR Langugage Technology Research Labor. "එකමුතුව යනු විවිධත්ව 🛁 '' එක මුතු වය නු වි වි ධ-ඒකාබද්ධතාවේ විශ්ලේෂණයෙන් ත්වයේ එක මූ තු වයි. එකමුතුවෙහි වටිතාකමට සමාත වි වි ධත්ව ය ව තා ඒකා මතු වී කිුයාත්මක වන්නේ විවිර බ ද්ධ) ත ා වේ වි ශ්ලේෂ කලා කෘතියේ එකමුතුව මනුෂා ණයෙන් පහළවෙයි <u>E</u>xit ඉතා මත් ම යෝගා වෙයි. කල . වි වි ධත්ව යෙ හි එක මූ ස්වයං-ශක්ති සම්පන්න ඒකීය තුවෙහිවටිතාකමට පවත්තේ විවිධ ඉන්දියන්ගේ සමානවෙයි. කුමක් සමස්තයෙන් වෙන් කළ කල්හි හෙයින් මතු වී කියා න් මේ සම්බන්ධයෙන් සුසාන් ක වන්නේ විවිධ කො ටස් පාලනයකිරීමෙ •

Figure 3.5.2

3.6 View Characters

In order to view the segmented characters click the "View Characters" button. Then the input box appears and inserts the line number, as shown in Figure 3.6.1

OCR Pre Proce	essor Ver0.1
(increased in the second secon	OCR PREPROCESSOR Langugage Technology Research Laboratory
<u>O</u> pen	Remove NoiseSegment LinesView LinesSegment CharactersView
	"
	''එකමුතුව යනු විවිධන්වයේ එකමුතු:
'' లి)කමුතුවයනු විවිධත්වයේ එක
	වයි. විවිධත්වය වතා ාහි
	Line No. 1

Figure 3.6.1: View Characters

3.7 Finding the properties of an Image

Click "Properties" button. Then the properties of the image appear in a massage box as shown in the Figure 3.7.1.

Open Noise Noise Segment Lines Segment Lines Segment Characters Properties Next Line Next Line PreprocessorVer1.0 Line Count = 37 Width = 888 Height = 1457	OCR Pre Processor V	'er0.1	×		
□pen Noise Remover Segment Lines Segment Next Line Characters Litopettes PreprocessorVer1.0 Next Character Exit Une Count = 37 Width = 888 Height = 1457		1000			
Line Count = 37 Width = 888 Height = 1457	linen i i	Ioise Segment Lines Segment <u>Characters</u> I			
Line No. 7Character No. 0		Line Count = 37 Width = 888 Height = 1457			

Figure 3.7.1: Properties of the image

3.8 Close the program

Click "Exit" button to close the application