

# Integration of Bangla script recognition support in OCRopus

By

Muttakinur Rahman Chowdhury (Shouro)

Supervisor

Dr. Mumit Khan

Co-supervisor

Md. Abul Hasnat

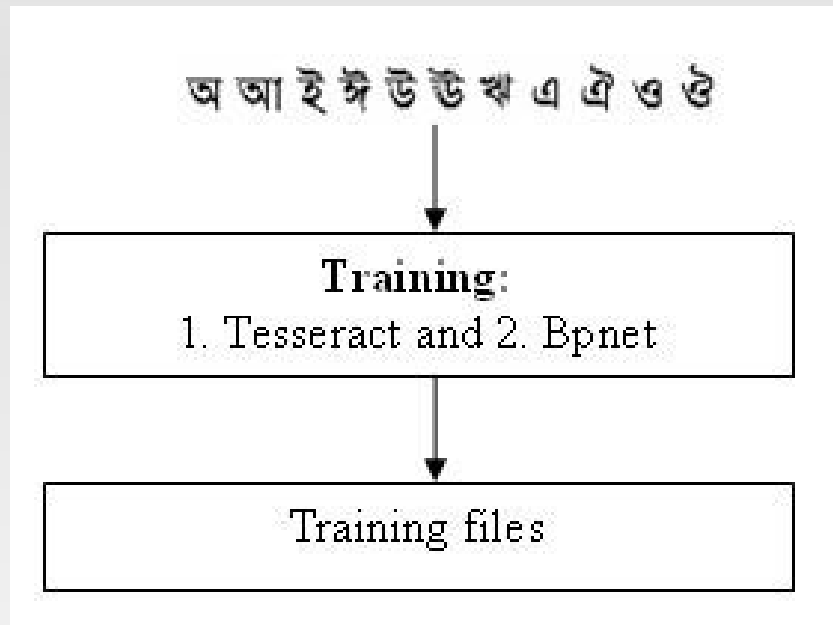
# What is OCRopus?

- OCRopus is an open source state-of-the-art document analysis and OCR system, featuring pluggable layout analysis, pluggable character recognition, statistical natural language modeling, and multi-lingual capabilities.

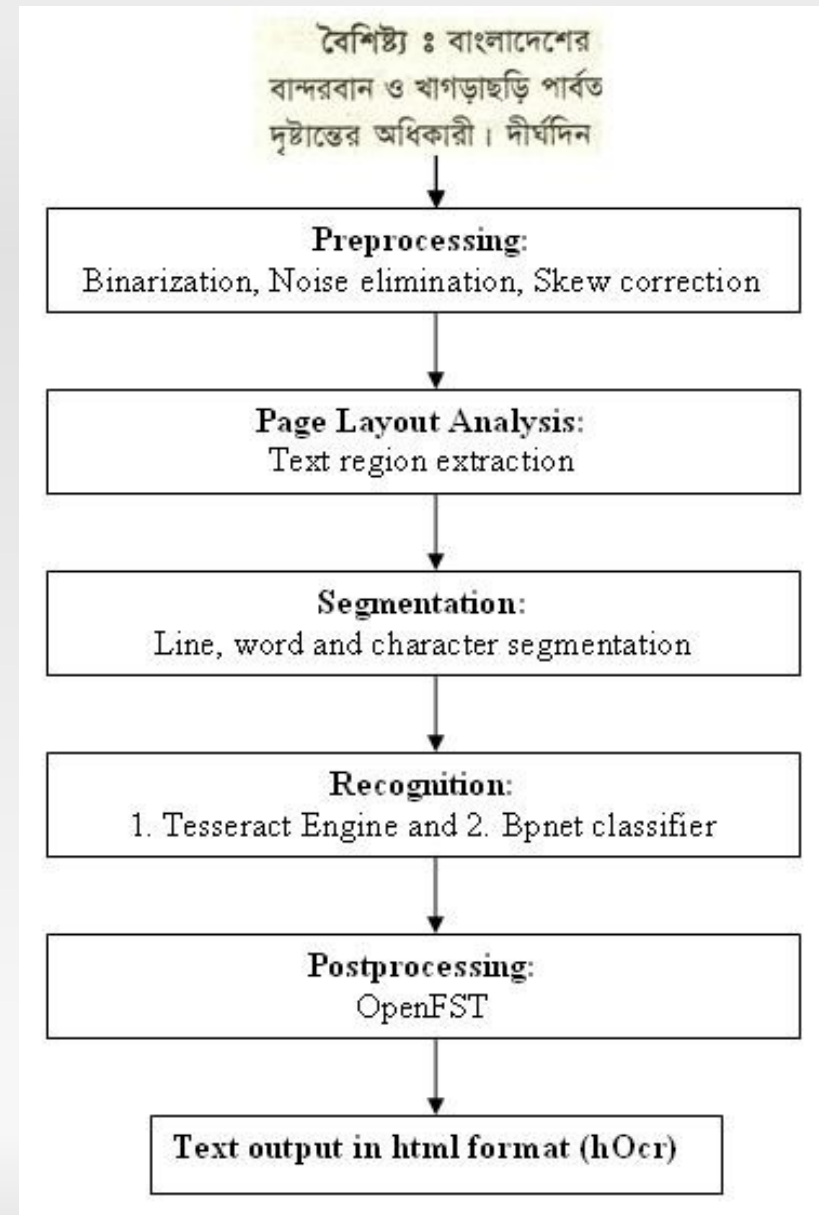
# Why OCRopus?

- A complete OCR framework.
- Multi-lingual capabilities.
- Rich image preprocessing and layout analysis library.
- Two recognition engines:
  - Tesseract.
  - bpnet (Multi Layer Perceptron Neural Network) classifier.
- Statistical natural language model (OpenFST) as a post processor.
- Developed with the generous support from Google and other organizations.

# Framework overview



## Training



## Recognition

# Training tesseract

- Step 1: Create training image data
- Step 2: Make Box file
- Step 3: Run Tesseract for Training
- Step 4: Clustering
- Step 5: Compute the Character Set
- Step 6: Prepare Dictionary data files

# Recognition tesseract

- Generate a properly segmented character image.
- Pass this image to the tesseract recognizer.

# Training bpnet

- Two Approaches:
  - From individual character image.
  - From text line image.
- 1st approach:
  - Individual character image.
  - Character unicode.
- 2nd approach:
  - Text line image.
  - Segmented color image of the text line image.
  - Transcription of the text image.

# Recognition bpnet

- Generate a color segmented character image.
- Pass this image to the bpnet classifier.

# Example:

## Training data generation for Tesseract

Step 1: Create training image data

```
ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ ণ ত থ দ ধ ন প ফ ব ভ ম য র ল শ  
ষ স হ ড় ঢ় য় ঞ  
অ ই ঙ্গ উ ঊ ঋ ঌ ঐ ও ঔ ০ ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯  
। ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯
```

Step 2: Make Box file

```
ক 32 306 82 354  
খ 102 308 143 363  
গ 164 306 205 363  
ঘ 221 307 265 354  
ঙ 283 305 321 350  
চ 341 304 379 348
```

Step 3: Run Tesseract for Training

```
tesseract trainfile.tif junk nobatch box.train
```

Step 4: Clustering

```
mftraining trainfile_1.tr trainfile_2.tr ...  
cntraining trainfile_1.tr trainfile_2.tr ...
```

# Example:

## Training data generation for Tesseract

Step 5: Compute the Character Set

```
unicharset_extractor trainfile.box
```

```
৩ 5 NULL  
ঔ 5 NULL  
০ 8 NULL  
৓ 8 NULL
```

Step 6: Prepare Dictionary data files

```
wordlist2dawg frequent_words_list freq-dawg  
wordlist2dawg words_list word-dawg
```

**Word List : 180K words**

**Frequent word List : 30K words**

1	কা	2	ক গ
2	ক গ	1	কা
2	তা	1	অ

**Reference** <http://crblpocr.blogspot.com/2008/07/how-to-train-bangla-and-devanagari.html>

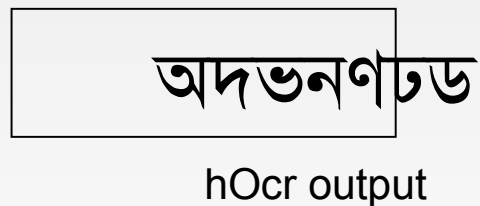
<http://crblpocr.blogspot.com/2008/08/why-tesseract-need-to-train-all.html>

# Example: Testing Tesseract

**Step 1:** Generate a properly segmented character image.



**Step 2:** Pass this image to the tesseract recognizer



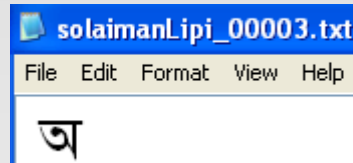
# Example:

## Training data generation for bpnet

### Approach 1: From individual character image



character image



Character unicode

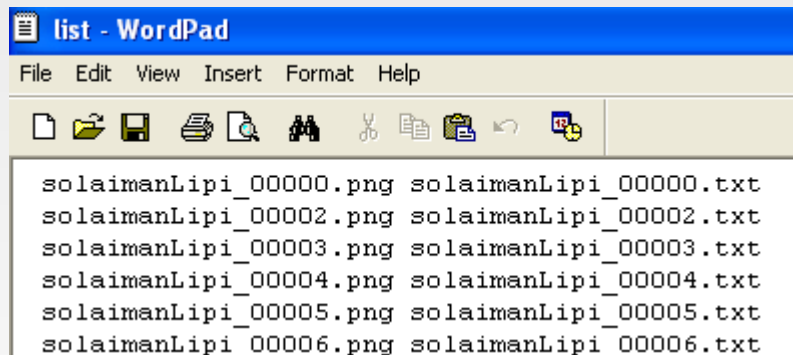


Image – transcription mapped list file

# Example:

## Training data generation for bpnet

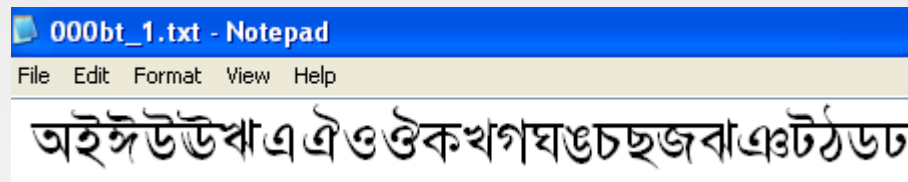
### Approach 2: From line image

অ ই ঙ উ ঊ ঋ এ ঐ ও ঔ ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ

Line image

অ আ ই ঙ উ ঊ ঋ এ ঐ ও ঔ ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ

Color image



Line transcription

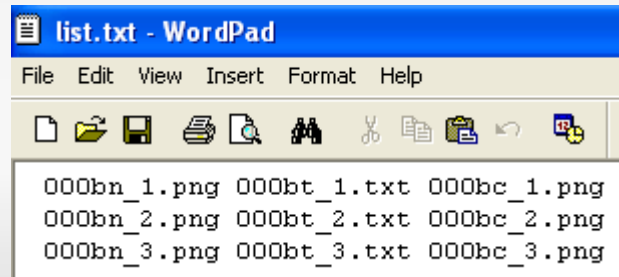


Image – transcription mapped list file

# Example:

## Testing bpnet classifier

**Step 1:** Generate a color segmented character image.

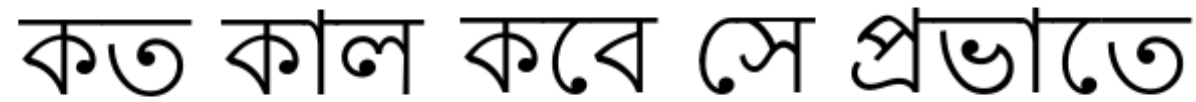
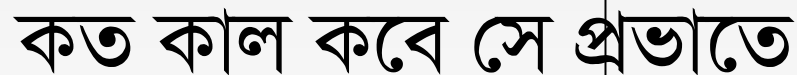
A white rectangular box containing the Bengali text "কত কাল কবে সে প্রভাতে" in black font.

Fig: Input text Image

A dark red rectangular box containing the Bengali text "কত কাল কবে সে প্রভাতে" where each character is a different color: ক (green), ত (blue), ক (red), ব (yellow),ে (purple), সে (pink), প্র (orange), ভ (light blue),া (light green), তে (green).

Fig: Segmented color Image

**Step 2:** Pass this image to the bpnet classifier.

A white rectangular box containing the Bengali text "কত কাল কবে সে প্রভাতে" in black font, identical to the input text image.

hOcr output

# Example:

## Testing bpnet classifier (best NN Parameters)

nhidden - 500  
epochs - 300  
learningrate - 0.2  
testportion - 0  
normalize - 1  
shuffle - 1

# CRBLP segmenter plugin

- Four step character segmentation algorithm.
  - Projection profile based.
  - Elimination of the joining errors
    - Shadow character segmentation
    - Touching character segmentation
  - Elimination of splitting errors
  - Reordering the modifiers

# Acknowledgement

- Mark Stillwell (University of Hawaii)
- Prof. Thomas Breuel (IUPR)
- Ilya Mezhirov (IUPR)
- Yves Rangoni (IUPR)
- Ray Smith (Google research)
  
- CRBLP Lab

# References

- <http://code.google.com/p/ocropus/>
- <http://code.google.com/p/tesseract-ocr/>
- <http://code.google.com/p/ocropus-bengali/>
- <http://crblpocr.blogspot.com/>

