

Indic Meet 2009, Pune

Free Indic OCR

Debayan Banerjee

NIT Durgapur

debayanin@gmail.com

Indic Meet 2009, Pune

What is OCR?

জন গণ মন অধি নায়ক জয় হে
ভারত ভাণ্ড্য বিধাতা ।

OCR

জন গন মন অধিনায়ক জয় হে
ভারত ভাণ্ড্য বিধাতা ।

Indic Meet 2009, Pune

Present Approach

- Modify Tesseract-OCR. Among the top 3 OCR software available for Latin languages. It is FOSS.

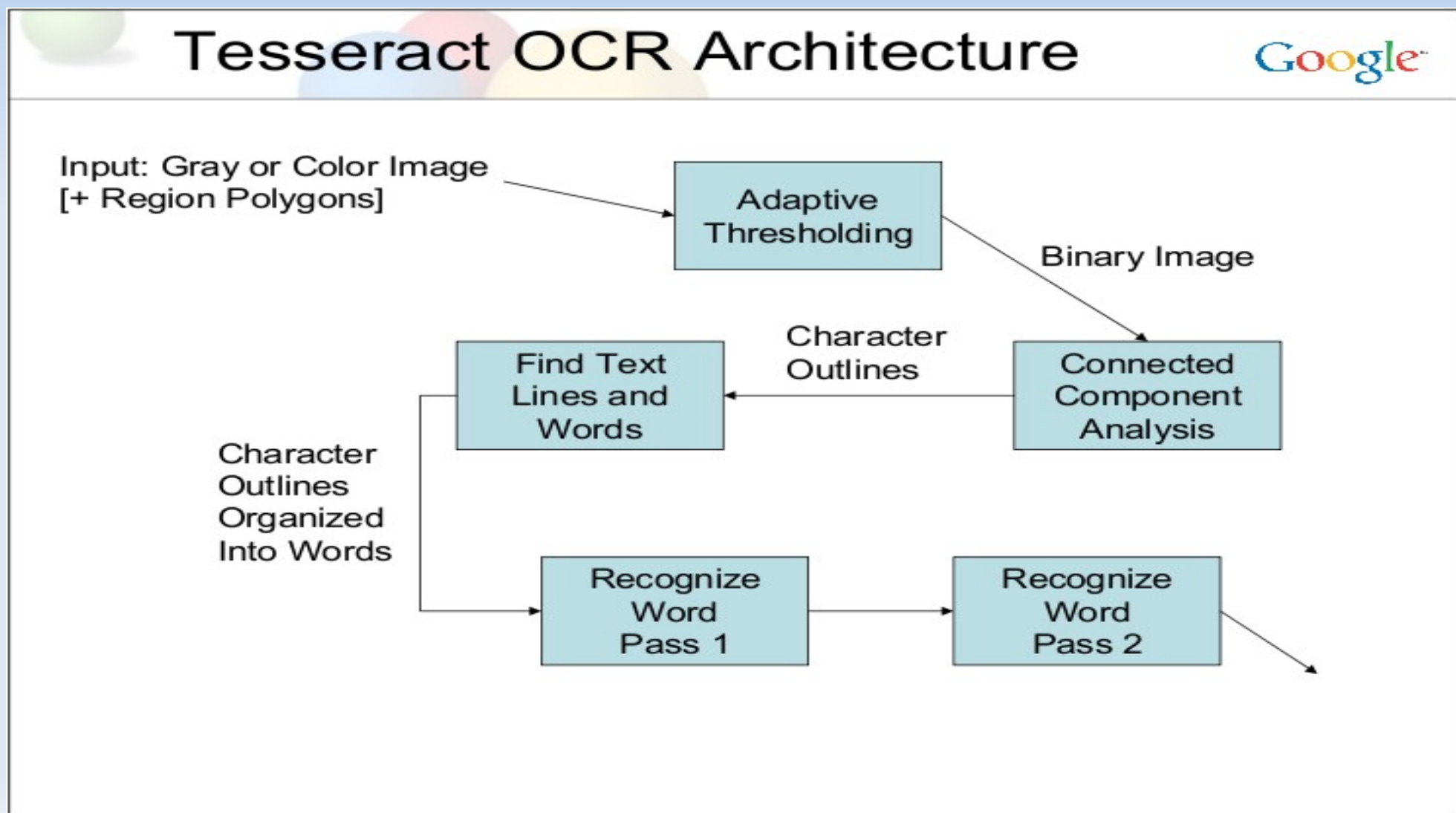
Indic Meet 2009, Pune

A little more information about Tesseract

- Developed on HP-UX at HP between 1985 and 1994 to run in a desktop scanner.
- Came neck and neck with Caere and XIS in the 1995 UNLV test. (See <http://www.isri.unlv.edu/downloads/AT-1995.pdf>)
- Never used in an HP product.
- Open sourced in 2005. Now on:
<http://code.google.com/p/tesseract-ocr>
- Highly portable. Search for Tesseract OCR on YouTube.

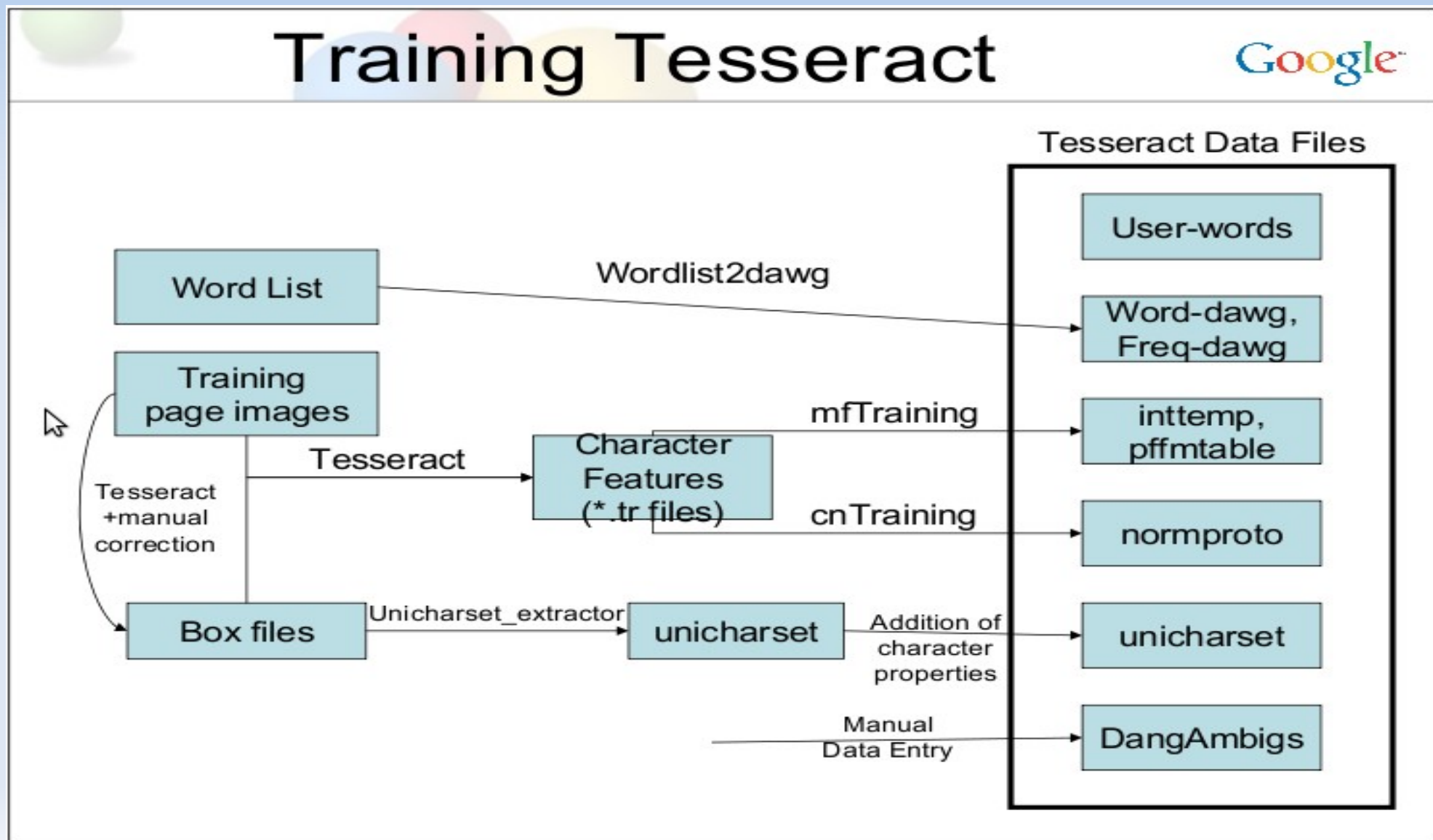
Indic Meet 2009, Pune

How does Tesseract-OCR work?



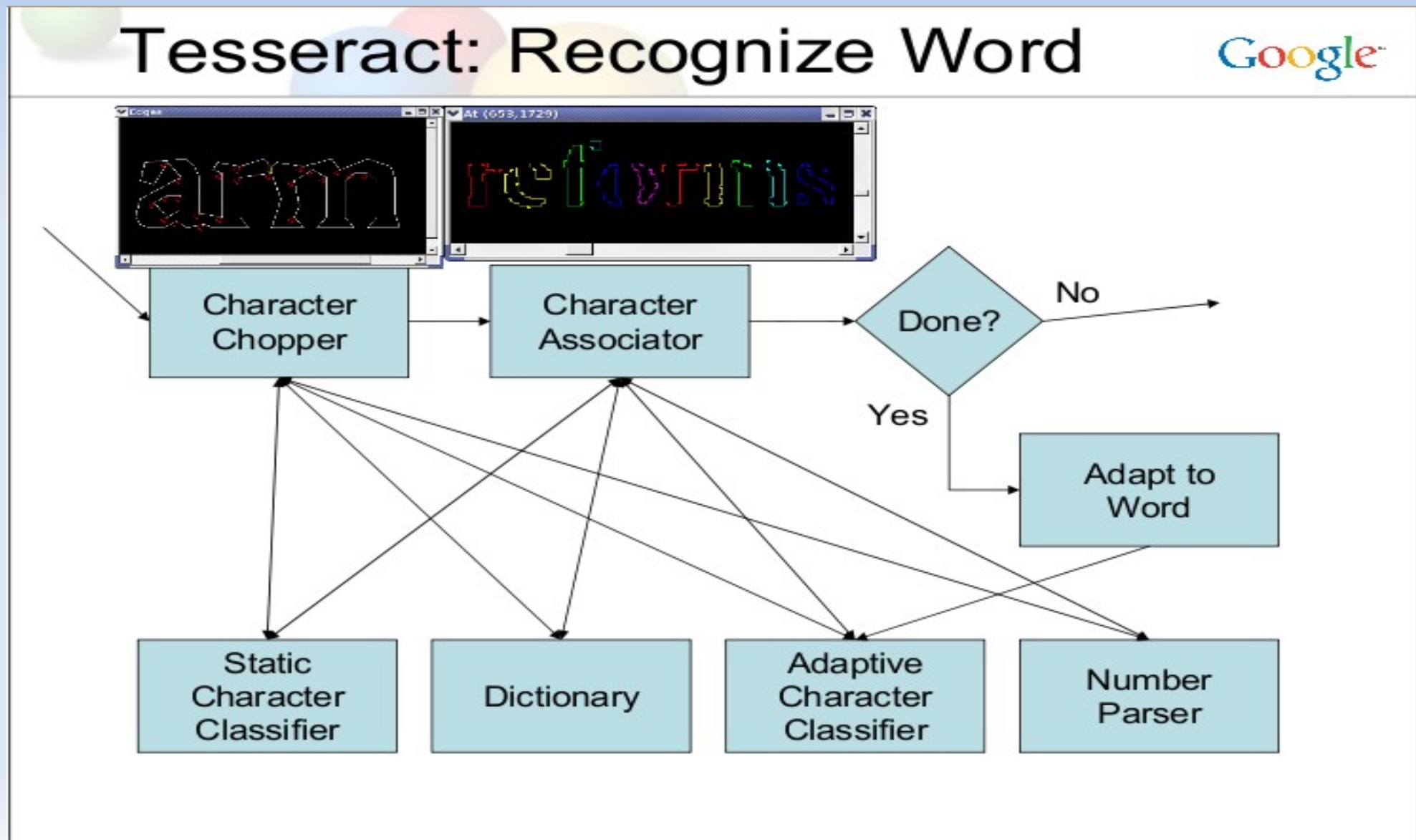
Indic Meet 2009, Pune

How does Tesseract-OCR work?



Indic Meet 2009, Pune

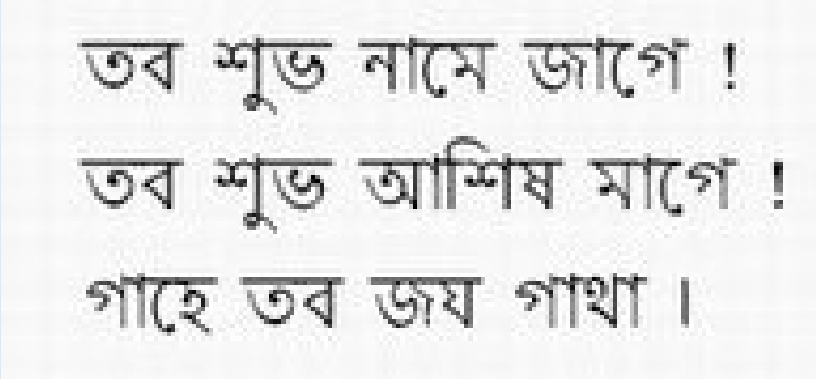
How Tesseract-OCR works?



Indic Meet 2009, Pune

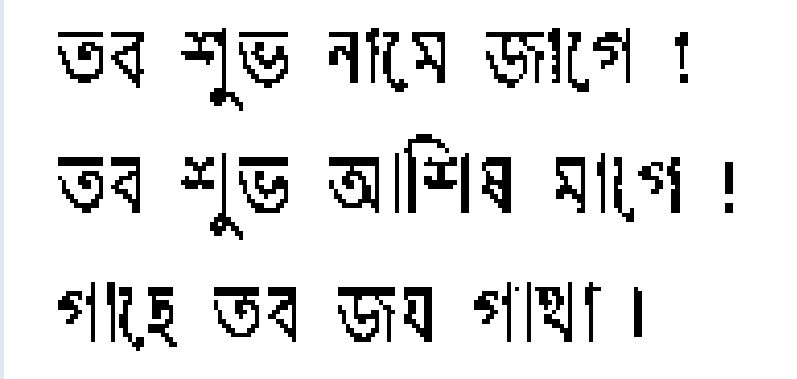
What parts do we need to hack to make it Indic capable?

1) Segmentation, Why?

The image shows three lines of Bengali text in a serif font, centered on a light beige background. The text is: "তব শুভ নামে জাগে !", "তব শুভ আশিষ মাগে !", and "গাহে তব জয় গাথা ।".

তব শুভ নামে জাগে !
তব শুভ আশিষ মাগে !
গাহে তব জয় গাথা ।

Original Image

The image shows the same three lines of Bengali text as the original, but rendered in black on a white background. The text is: "তব শুভ নামে জাগে !", "তব শুভ আশিষ মাগে !", and "গাহে তব জয় গাথা ।".

তব শুভ নামে জাগে !
তব শুভ আশিষ মাগে !
গাহে তব জয় গাথা ।

Thresholded and clipped image

Indic Meet 2009, Pune

2) Orientation. Deskew. Why?

Images and Figures

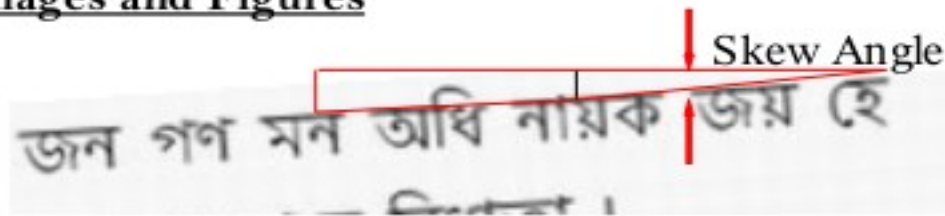


Fig 1

জন গণ মন অধি নায়ক জয় হে
ভারত ভাষ্য বিধাতা ।
পঞ্জাব সিন্ধু গুজরাত মরাঠা
দ্রাবিড উৎকল বঙ্গা ।
বিংগ্ব হিমাচল যমুনা গঙ্গা
উচ্ছল জলধি তরংগা ।
তব শুভ নামে জাগে !
তব শুভ আশিষ মাগে !
গাহে তব জয় গাথা ।
জন গণ মঙ্গলদায়ক জয় হে
ভারত ভাষ্য বিধাতা ।
জয় হে ! জয় হে ! জয় হে !
জয় জয় জয় জয় হে !

Fig 2

Tilted scanned image

Indic Meet 2009, Pune

Orientation continued....

জন গণ মন অধি নায়ক জয় হে
ভারত ভাষ্য বিধাতা ।
পঞ্জাব সিন্ধু গুজরাত মরাঠা
দ্রাবিড় উৎকল বঙ্গা ।
বিংশয় হিমাচল যমুনা গঙ্গা
উচ্ছল জলধি তরংগা ।
তব শূভ নামে জাগে !
তব শূভ আশিষ মাগে !
গাহে তব জয় গাথা ।
জন গণ মঙ্গলদায়ক জয় হে
ভারত ভাষ্য বিধাতা ।
জয় হে ! জয় হে ! জয় হে !
জয় জয় জয় জয় হে !

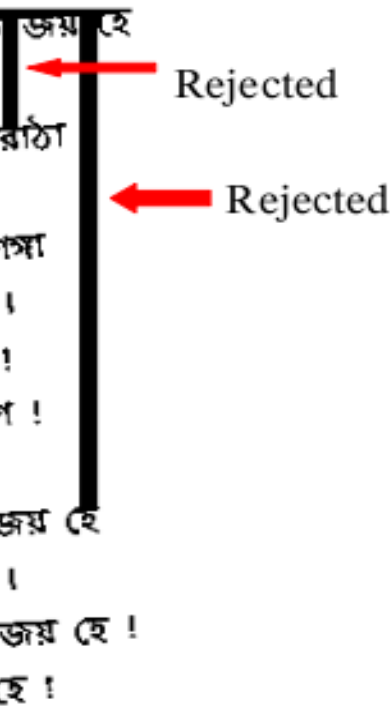


Fig 3

This image has been generated by tracing the execution of the algorithm by using black pixels. The long black pixels have been rejected for angle calculation because the slope they give us are too large (>10 degrees).

জন গণ মন অধি নায়ক জয় হে
ভারত ভাষ্য বিধাতা ।
পঞ্জাব সিন্ধু গুজরাত মরাঠা
দ্রাবিড় উৎকল বঙ্গা ।
বিংশয় হিমাচল যমুনা গঙ্গা
উচ্ছল জলধি তরংগা ।
তব শূভ নামে জাগে !
তব শূভ আশিষ মাগে !
গাহে তব জয় গাথা ।
জন গণ মঙ্গলদায়ক জয় হে
ভারত ভাষ্য বিধাতা ।
জয় হে ! জয় হে ! জয় হে !
জয় জয় জয় জয় হে !

Indic Meet 2009, Pune

- Algorithm available at http://tesseractindic.googlecode.com/files/skew_deskew.pdf
- Better deskewing and pre-processing methods available. No need to worry.
- Ocropus

Indic Meet 2009, Pune

3) Training. Why? (Ans: Tedious as hell!)

Take a list of characters → Render them.
Introduce degradation to get more samples →
Find out their bounding boxes → Tell
Tesseract-OCR → It generates training data
files

Done using Python scripts. PIL and Pango.

Indic Meet 2009, Pune

4) Spell-check. Why?

- What is DAWG? Why does it not work for Indic?
- Pyhunspell?
- We need to repair or replace DAWG probably.

Indic Meet 2009, Pune

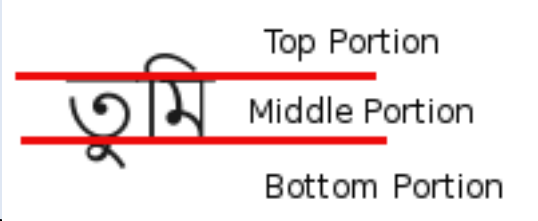
Demo time :)

Indic Meet 2009, Pune

Problem 1:

1) Under 50 trainable characters in English.
1800 trainable characters in Bengali/Hindi. Why?

Possible Solutions:

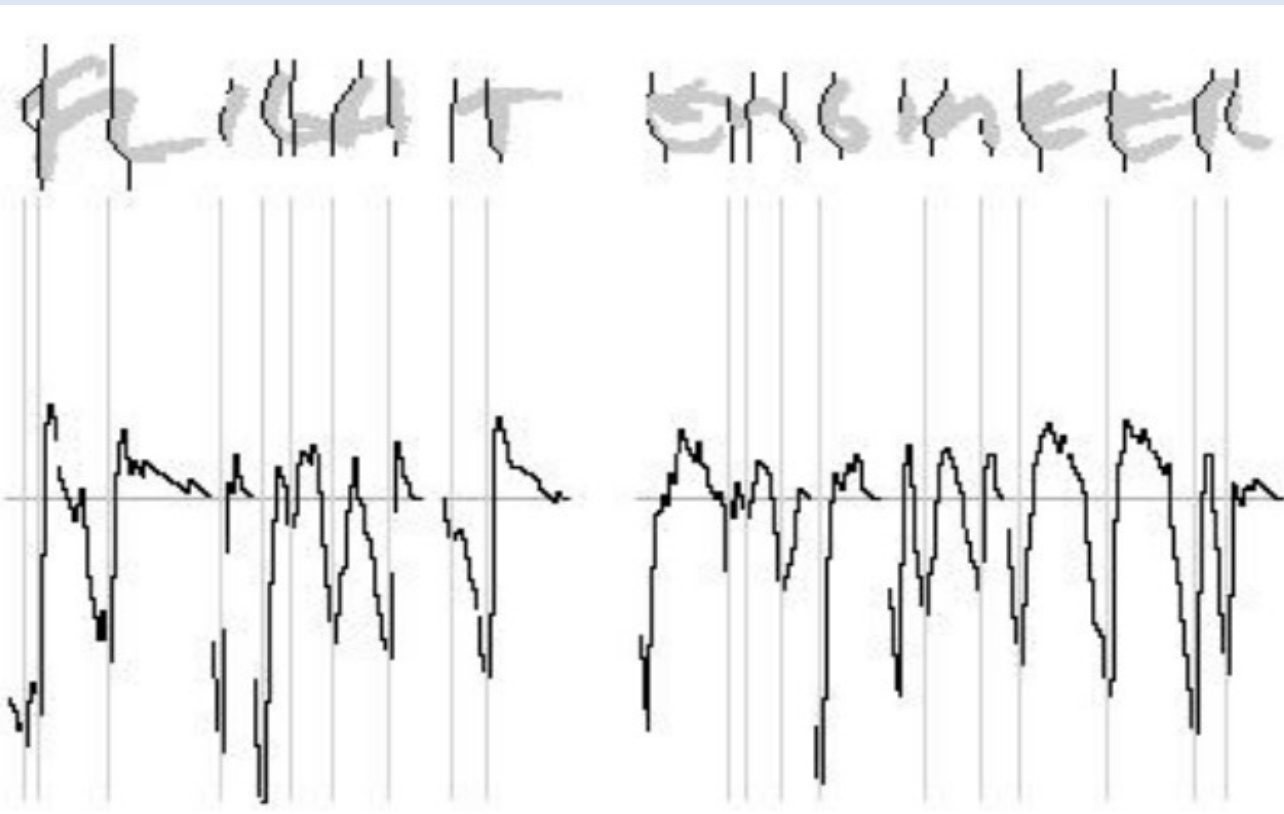
1)  Divide the image into zones.
Reduces character classes to
around 300. Lots of hacking to be done in
Tesseract-OCR to achieve this. Why?

Indic Meet 2009, Pune

Segmentation continued...

2) Or we could use curved cut segmenters.

॥ स्वस्ति श्रीश¹पदपङ्कजसेवनावाप्तिसकलमनोरथानां सह-



Ocropus Curved Cut Segmenter at Work

Indic Meet 2009, Pune

Problem 2:

Make the OCR available to public easily.

Solution:

- Web Interface. Shantanu from Sarai developed one for me. I developed some more.
- Need to integrate with Silpa. Santhosh?
- Create a feedback based learning system. Do-able with Python scripts.

Indic Meet 2009, Pune

Problem 3:

Need ample testing data. That means images with corresponding text, with each word on a separate line.

Solution:

- Professor from ISI Kolkata provided me with some.

Indic Meet 2009, Pune

Problem 4:

Finding new project members. Learning curve is steep. Documentation is very important.

Solution:

- Many of my juniors work at ISI Kolkata in the summers in the Computer Vision Department.
- Google keeps telling us that they are working on an Indic OCR. Its been over a year though.

Indic Meet 2009, Pune

Opportunities:

- Many people work on Digital Image Processing in the summers. OCR is a common topic
- Jinesh K J and Indu S are working with me on this
- Technical support available in terms of Professors and papers.
- In parts, the problem is already solved. Some more hacking involved.

Indic Meet 2009, Pune

Target:

- 98% accuracy for a given font and size
- Rapid/trivial training data generation
- Given ground truth data, accuracy calculation must be automated
- Learn from web based feedback
- Develop an interface for the same

Indic Meet 2009, Pune

Thanks :)

Debayan Banerjee
DeepRoot Linux, <http://deeproot.in>
debayanin@gmail.com