State-of-the-art OCR technology

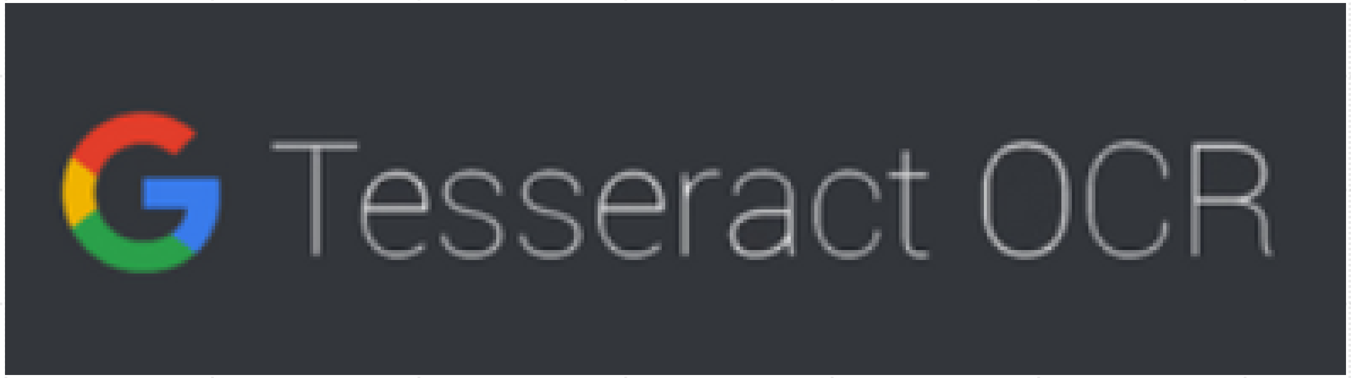
- Mihir Shah

Agenda

- State-of-the-art Tesseract OCR technology
- Architecture
- Training Tesseract
- Evaluation

Tesseract OCR technology

- Open-source OCR engine
- Developed and Maintained by Google
- Can detect over 100 languages
- Can process even right-to-left text



Architecture Tesseract 4.0

- New OCR engine that uses LSTM (Long Short-Term Memory)
- Most effective solution for sequence prediction problems
- Pytesseract python wrapper

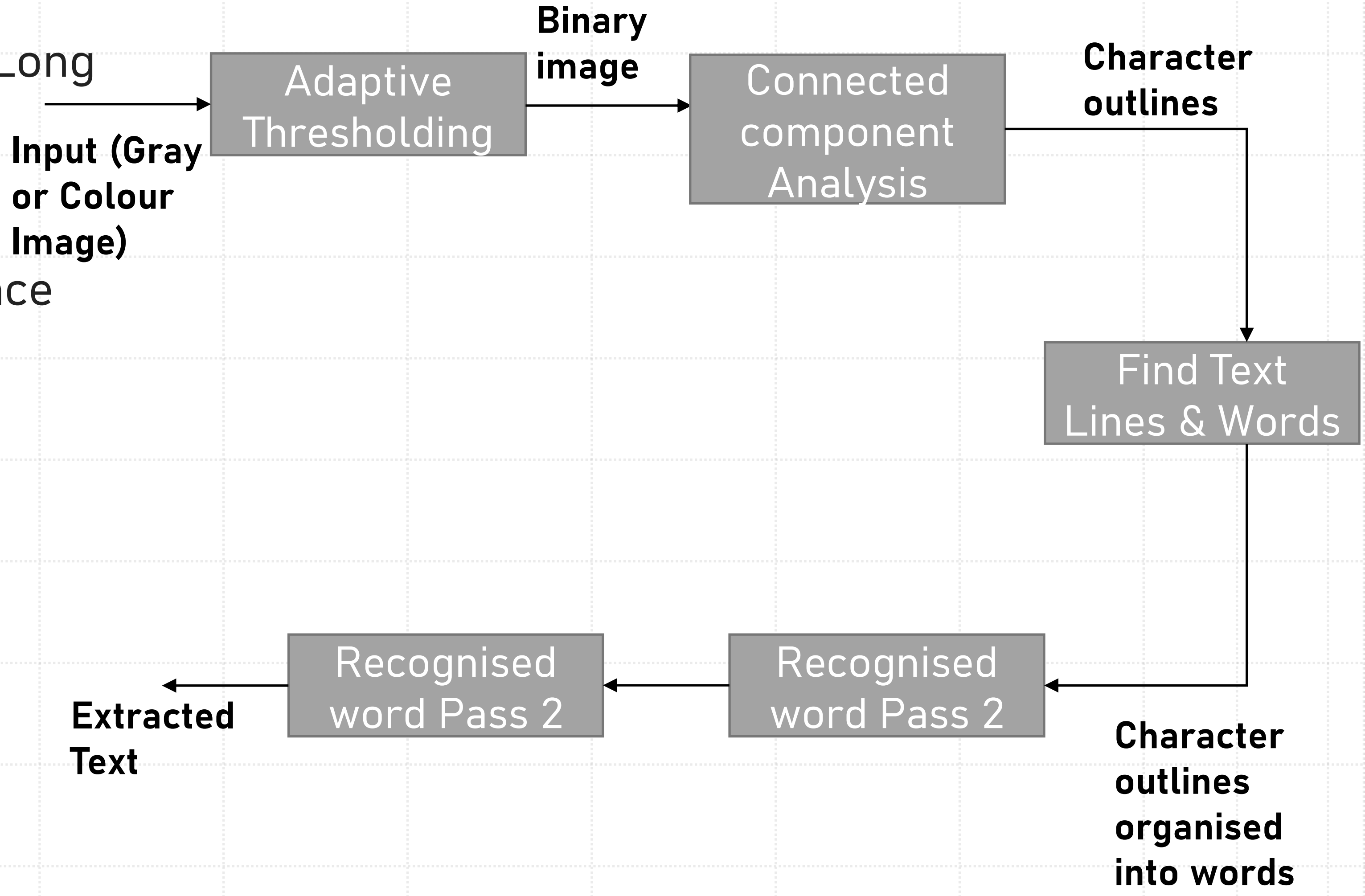


Figure: Architecture of Tesseract 4.0

Training Tesseract 4.0

- Training for custom fonts
- LSTM Model

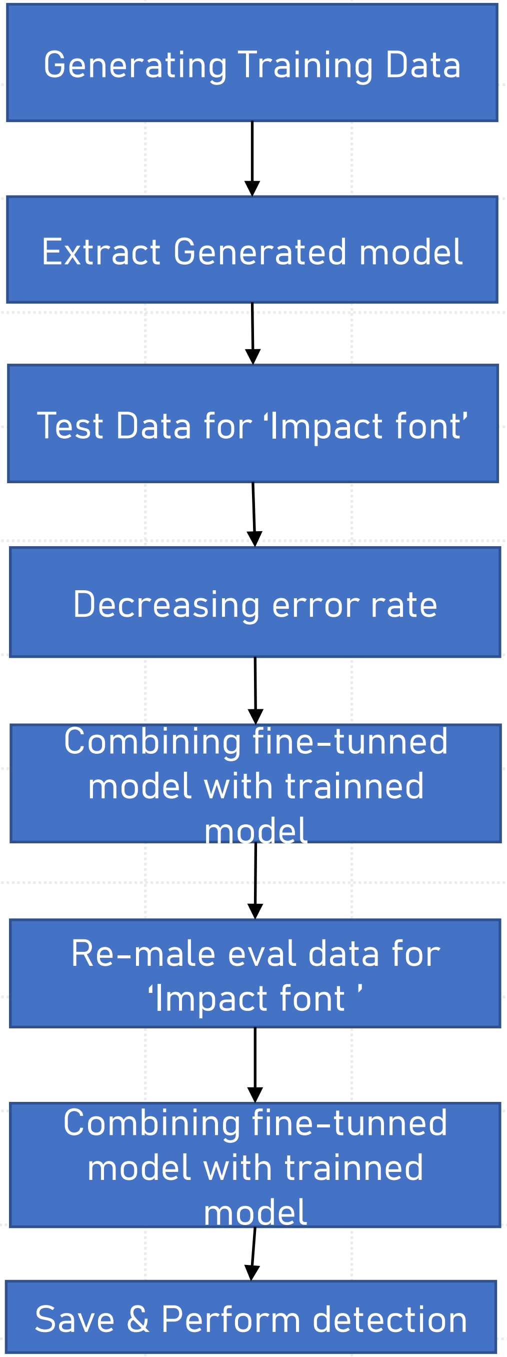

Generating Training Data
Extract Generated model
Test Data for 'Impact font'
Decreasing error rate
Combining fine-tunned model with trainned model
Re-male eval data for 'Impact font '
Combining fine-tunned model with trainned model
Save & Perform detection

Figure: Flow Chart for Training Tesseract 4.0

Evaluation Metrics

- **Character Error Rate** (CER) : Represents the **percentage** of characters that were **incorrectly** predicted.
- **Word Error Rate** (WER): Computes the **minimum edit distance** between a human-generated sentence and the machine-predicted sentence.

$$\text{WER} = \frac{Number\ of\ Errors}{Totel\ Words}$$

Thank you!