Tibetan OCR Test Corpus Summary

Overview

Location: `~/Documents/tibetan-ocr-app/input_files/tibetan_texts/`

Total PDFs: 33 files organized by script type and quality level

---

Script Types & OCR Models

UCHEN (Block Print Script)

OCR Models to Test:

- `Woodblock`
- `Woodblock-Stacks`
- `Modern`

Test Files by Quality:

High Quality (3 files)

Pages	Size	File
10	192K	`uchen high quality pdf.pdf`
3	588K	`uchen_high_quality_.pdf`
3	900K	`uchen_high_quality_pdf.pdf`

Medium Quality (2 files)

Pages	Size	File
3	1.6M	`uchen_medium quality.pdf`
4	2.0M	`uchen_medium quality(1).pdf`

Poor Quality (3 files)

Pages	Size	File
2	252K	uchen_tsalyig_poor quality.pdf
3	1.6M	uchen_poor quality.pdf
4	180K	uchen_poor quality(1).pdf

Total Uchen Files: 8 files, 31 pages

UMEH (Cursive Script)

OCR Models to Test:

- Ume_Druma
- Ume_Petsuk
- Modern

Umeh Variants Represented:

- Dhernangdri (most common in test set)
- Druchen
- Drutsa
- Tsugma/Khyug
- Petsug
- Khyugyig

Test Files by Quality:

High Quality (6 files)

Pages	Size	Variant	File
3	548K	Druma	umeh_druma_high quality.pdf
3	428K	Drutsa	umeh_drutsa_high quality.pdf
3	592K	Dhernangdri	umeh_dhernangdri_high quality(1).pdf
3	704K	Dhernangdri	umeh_dhernangdri_high quality.pdf
3	856K	Dhernangdri	umeh_dhernangdri_high quality(2).pdf
3	680K	Tsugma Khyug	umeh_tsugma khyug_high quality.pdf

Medium Quality (5 files)

Pages	Size	Variant	File
3	480K	Druchen	umeh_druchen_medium quality.pdf
4	2.0M	Druchen	umeh_druchen_medium quality(1).pdf
12	980K	Dhernangdri	umeh_dhernangdri_medium quality.pdf
3	696K	Dhernangdri	umeh_dhernangdri_medium quality(1).pdf
3	784K	Dhernangdri	umeh_dhernangdri_medium quality(2).pdf

Poor Quality (6 files)

Pages	Size	Variant	File
3	368K	Dhernangdri	umeh_dhernangdri_poor quality.pdf
3	140K	Dhernangdri	umeh_dhernangdri_poor quality(1).pdf
3	784K	Dhernangdri	umeh_dhernangdri_poor quality(2).pdf
3	416K	Drutsa	umeh_drutsa_poor quality.pdf
3	112K	Khyugyig	umeh_khyugyig_poor quality.pdf
3	308K	Petsug	umeh_petsug_poor quality.pdf

Total Umeh Files: 17 files, 61 pages

Additional Files (Script Type TBD)

Pechas with More Text (3 files)

Pages	Size	File
3	1.8M	pechas with more text 1.pdf
3	1.6M	pechas with more text 2.pdf
3	1.4M	pechas with more text 3.pdf

Pechas with Little Text (3 files)

Pages	Size	File
3	2.2M	pechas with little text 1.pdf
3	2.1M	pechas with little text 2.pdf
3	1.2M	pechas with little text 3.pdf

Standalone Files (2 files)

Pages	Size	File
2	72K	sangs rgyas sman bla dngos.pdf
12	5.5M	sangs_rgyas_sman_gyi.pdf

Grid Search Parameters

Parameters Per Image

Total Combinations: 1,728 per image

Parameters:

- **OCR Models:** 3 models (varies by script type)
- **line_mode:** 2 options (line, layout)
- **class_threshold:** 3 values (0.7, 0.8, 0.9)

- **k_factor:** 3 values (2.0, 2.5, 3.0)
- **bbox_tolerance:** 4 values (2.5, 3.5, 4.0, 5.0)
- **merge_lines:** 2 values (True, False)
- **tps_threshold:** 4 values (0.1, 0.25, 0.5, 0.9)

Calculation: 3 × 2 × 3 × 3 × 4 × 2 × 4 = 1,728

First Test Recommendation

File: `sangs rgyas sman bla dngos.pdf`

- **Pages:** 2 (smallest in collection)
- **Size:** 72K
- **Total OCR runs:** 2 pages × 1,728 = 3,456 runs
- **Estimated time:** 30-45 minutes
- **Script type:** Need to verify (likely Uchen based on filename)

Location for test:

```
bash

~/Documents/tibetan-ocr-app/test_samples/first_test/
```

Quality Scoring

Each OCR output will be automatically scored using **PyBo tokenization**:

- **Score:** 0-100 (percentage of valid Tibetan words)
- **Threshold guidance:**
  - > 90: Excellent OCR
  - 70-90: Good OCR
  - 50-70: Fair OCR
  - <50: Poor OCR (likely garbage)

Results will be saved in `summary.csv` sorted by quality score, allowing focus on top-performing parameter combinations.

Total Corpus Statistics

Category	Files	Total Pages
Uchen High	3	16
Uchen Medium	2	7
Uchen Poor	3	9
Umeh High	6	18
Umeh Medium	5	25
Umeh Poor	6	18
More Text	3	9
Little Text	3	9
Standalone	2	14
TOTAL	**33**	**125**

If all files processed:

- Total images: ~125 pages
- Total OCR runs: 125 × 1,728 = **216,000 runs**
- Estimated time: 50-80 hours (depends on machine)