Scene Text Data Collection & Evaluation Using IndicPhotoOCR

What We Did

Used the existing IndicPhotoOCR toolkit to test scene-text recognition.

Collected real-world images containing Indian-language text.

Manually wrote (annotated) the visible text in each image.

Compared our annotations with the OCR output to check performance.

Goal:

Understand how a basic ML pipeline works by using an existing OCR system.

Contributors

- Tirunagari Bhuvan Sri Sai (B24BB1040)
- Kesanapalli Jithin (B24CM1037)
- Jeram Arjun (B24CM1034)
- Gopi Sumanth Chadalavada (B24CM1084)

MOTIVATION

Text in natural scenes (shop boards, buses, streets) is common but hard for OCR.

Indian scripts are challenging because of many languages, curved shapes, and clumsy backgrounds.

IndicPhotoOCR already supports multiple Indian languages.

By collecting and annotating images, we help improve evaluation of Indian-language OCR tools.

Our Tasks

1. Data Collection

Collected 50 images from Jodhpur/IITJ or hometown areas.

Ensured variety: different lighting, angles, backgrounds, and languages.

2. Annotation

For each image, we typed exactly what text appears.

Saved the annotations in simple text files or JSON.

3. Evaluation (Using IndicPhotoOCR)

Ran the existing IndicPhotoOCR model on our dataset.

Compared OCR predictions with our manually written text.

Noted where the model worked well and where it failed.

Workflow & Tools Used

•**IndicPhotoOCR:** used for text detection + recognition.
•**GitHub:** used for storing images, annotations, scripts, and results.
•Followed a simple workflow:
Collect Images → Annotate Text → Run OCR → Compare Results → Upload to GitHub
•Learned about:
•How OCR systems behave on real images
•Dataset preparation
•Evaluation basics
•Collaboration through GitHub

Observations

Highly trained(exposed languages) were having better accuracy while using the indic-photo-ocr while the less exposed (trained) data set languages , due to less exposure didn’t have accuracy close to those of the highly trained languages.

While sometimes , less exposed languages like Telugu had sometimes shown good results proportional to clarity, but as quality of image decreased slightly accuracy impacted significantly.

Lack of training data has worsened the accuracy for the less trained languages.

Results & Learnings

Key Observations

OCR worked well on clear, front-facing text.

Struggled with:
Shadows / low lighting
Angled or curved text
Highly stylized fonts
Crowded backgrounds

What We Learned

How to prepare real-world data for ML evaluation.

Importance of clean annotations for testing a model.

How to use tools like GitHub and pull requests.

Understanding strengths and weaknesses of existing OCR systems.