EEE6778 – Applied Machine Learning II (Technical Blueprint Report For Bilingual Handwriting Translator Project) - Venkata Jyothi Priya Mulaka

a. Problem Context and Project Summary

Handwriting is a personal and cultural form of visual expression. Yet, most AI translation systems today only translate *meaning* across languages, ignoring the *style* that defines how writing looks. There are no tools capable of rendering English text in the visual handwriting style of Telugu (and vice versa). This limits creativity for designers, educators, and students, and hinders digital preservation of handwritten scripts.

The goal of this project is to design a **Bilingual Handwriting Translator (English ↔ Telugu)** that converts handwritten or typed English text into Telugu handwriting style — and the reverse — using a **CycleGAN** model. Unlike traditional text translation, this system focuses on *style translation* while preserving semantic meaning. The project will also integrate an **EasyOCR-based validation module** to ensure text fidelity and provide an **interactive Streamlit interface** for visualizing the translation process.

b. Dataset

1. IAM Handwriting Dataset (English)

- **Source:** Kaggle – IAM Handwriting Word Database
- **Type:** Image dataset of English handwritten words.
- **Format:** `.png` images, 256 grayscale, 300 dpi.
- **Size:** ~1.6 GB (subset of 115,000 word samples).
- **Structure:** Word-level images organized in nested writer folders (e.g., `a01`, `b01`, etc.).
- **Access Method:** Downloaded directly from Kaggle and stored locally on `D:\datasets\IAM_English`.
- **Use:** Source domain for English handwriting style.

2. IIIT-HW Telugu Dataset (v1)

- **Source:** CVIT, IIIT Hyderabad
- **Type:** Telugu handwritten word images.
- **Format:** .jpg images, labeled using vocabulary IDs in accompanying train, val, and test files.
- **Size:** Approximately 3.6 GB, containing hundreds of thousands of word-level samples.

- **Structure:** Extracted from TeluguSeg.tar.gz into the folder path D:\datasets\IIIT-HW-Telugu_v1\TeluguSeg. The images are stored in multiple nested subfolders organized by writer IDs and page numbers.
- **Access Method:** Images are loaded recursively from all subfolders using the glob pattern ("**/*.jpg", recursive=True), without flattening the directory structure. This allows efficient file access and consistent data loading during training

3. Data Format and Preprocessing

Both datasets contain unpaired handwriting images. Since CycleGAN does not require paired samples, English and Telugu words are randomly matched during training.

Preprocessing Techniques (using OpenCV):

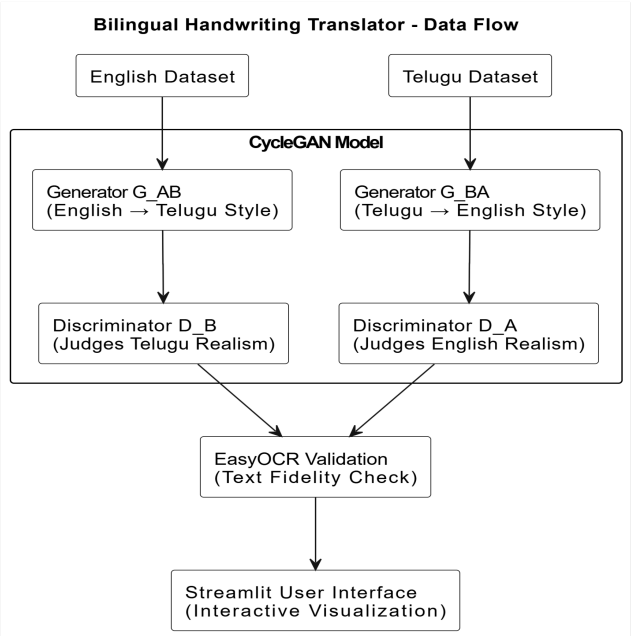
- **Resizing:** 128×128 pixels.
- **Grayscale conversion and normalization:** Range [-1, 1].
- **Dynamic augmentations during training:**
    - *Gaussian Blur:* adds cursive smoothness.
    - *Dilation:* mimics bold handwriting.
    - *Erosion:* creates lighter, calligraphic effects.
    - *Brightness/Contrast adjustment:* simulates ink and lighting variation. These augmentations increase stylistic diversity without enlarging dataset size.

4. Ethical or Privacy Considerations

- Both datasets are open-access and used solely for research.
- No personal or identifying data is included.
- Outputs are restricted to educational and creative use — not for signature imitation or identity replication.

c. Planned Architecture

Data Flow Diagram



Model Design

- **Core Model:** CycleGAN for unpaired image-to-image translation.
- **Framework:** PyTorch with CUDA acceleration (training on CPU initially; scalable to GPU).
- **Loss Functions:**
    - Adversarial Loss (LSGAN variant)
    - Cycle Consistency Loss (L1)
    - Identity Loss for handwriting texture preservation
- **Training Parameters:**
    - Image size = 128×128, batch size = 16
    - Optimizer: Adam (lr = 2e-4)
    - Normalization: InstanceNorm2D
- **Post-Training OCR Validation:**
    - EasyOCR extracts text from generated images.
    - If OCR text differs from the original, output is flagged for refinement.

Frameworks & Tools

- **Modeling:** PyTorch, Torchvision
- **Preprocessing:** OpenCV, NumPy, Pillow
- **OCR Validation:** EasyOCR
- **Interface:** Streamlit (interactive browser app)

d. User Interface Plan

Platform

A **Streamlit** web interface will make the project interactive and interpretable.

User Input

- Upload a handwritten English or Telugu image (JPEG/PNG).
- Or enter text to generate handwriting-style translation.
- Select translation direction (English → Telugu or Telugu → English).
- Choose a visual style filter: *Cursive*, *Bold*, or *Calligraphy*.

System Output

- **Original image** (user input).
- **OCR-detected text** displayed in typed font.
- **Generated translation** (CycleGAN output).
- **OCR Match Score** showing text-fidelity.
- Option to download the generated image.

Wireframe Sketch

Usability Impact

- Simple layout for quick comparison between source and translated handwriting.
- Integrated OCR transparency improves trust and interpretability.
- Supports educational and creative exploration for bilingual design.

e. Innovation and Anticipated Challenges

Innovative Aspects

- First project focused on **cross-script handwriting style translation** (English ↔ Telugu).
- **Integration of EasyOCR** for automatic text-fidelity verification.
- **Dynamic OpenCV-based augmentation pipeline** simulating natural handwriting variations.
- **Interactive Streamlit UI** that bridges technical deep learning outputs with user-friendly visualization.

Anticipated Challenges & Mitigation

Challenge	Description	Mitigation Strategy
1. Large dataset size (6+ GB)	CPU extraction/training slow	Use recursive data loading, subset sampling, and incremental training
2. CycleGAN instability	GANs prone to mode collapse	Add learning-rate scheduler, identity loss, and gradient clipping
3. Text preservation during translation	Handwriting deformation can reduce legibility	Integrate EasyOCR post-check and tune loss balance between adversarial & cycle losses

f. Implementation Timeline

Week	Focus	Expected Outcome
Oct 20 – Oct 26	Dataset extraction, preprocessing, setup notebook	Working dataset loaders for English & Telugu
Oct 27 – Nov 2	Baseline CycleGAN implementation	Basic style translation working

Nov 3 – Nov 9	Add OpenCV augmentations & improve stability	Realistic handwritten effects
Nov 10 – Nov 16	Integrate EasyOCR for validation	OCR pipeline verifying generated outputs
Nov 17 – Nov 30	Build Streamlit UI & link inference	Interactive bilingual handwriting translator
Dec 1 – Dec 11	Final optimization, testing & demo report	Completed end-to-end system & presentation

g. Responsible AI Reflection

Fairness:
The system equally supports two linguistic groups — English and Telugu — without favoring one. Future extensions could include more Indic languages for inclusivity.

Transparency:
Users see both OCR output and generated handwriting, ensuring visibility into what the model “understands” and how it transforms data.

Privacy:
All data are public handwriting samples; no personal or biometric information is processed.

Environmental Impact:
Training on large GANs consumes energy; to minimize this, only subsets will be trained on CPU with early stopping and reduced image sizes (128×128).

Note: The Streamlit interface design presented here serves as a preliminary wireframe for conceptual understanding. During implementation, the layout and interaction flow may be refined based on technical feasibility, model output behavior, or user experience improvements. Any necessary modifications or enhancements will be incorporated while maintaining the core functionality and interpretability goals of the project.