Milestone 1: Dataset Collection & Preprocessing Project: Trace Finder - Forensic Scanner Identification

1. Project Overview

The primary objective of this milestone was to identify the source scanner device used to scan a document by analyzing unique patterns or artifacts. This is achieved by capturing specific noise, textures, or compression traces introduced by the scanner hardware during the digitization process.

2. Data Collection (Week 1)

I have successfully collected scanned document samples from multiple devices.

- **Target Scanners:** A minimum of 3–5 scanner models/brands were identified.
- **Document Types:** The samples include various document types, such as Wikipedia pages and official documents, to provide a variety of textures for analysis.
- **Labeling:** A labeled dataset was created by assigning proper labels based on the source device and organizing them into specific folders.

3. Basic Image Analysis

An analysis of the raw samples was conducted to understand image properties:

- **Format:** Images are analyzed based on their format and color channels.
- **Resolution:** Basic properties like resolution were analyzed to ensure they are suitable for forensic extraction.

4. Image Preprocessing (Week 2)

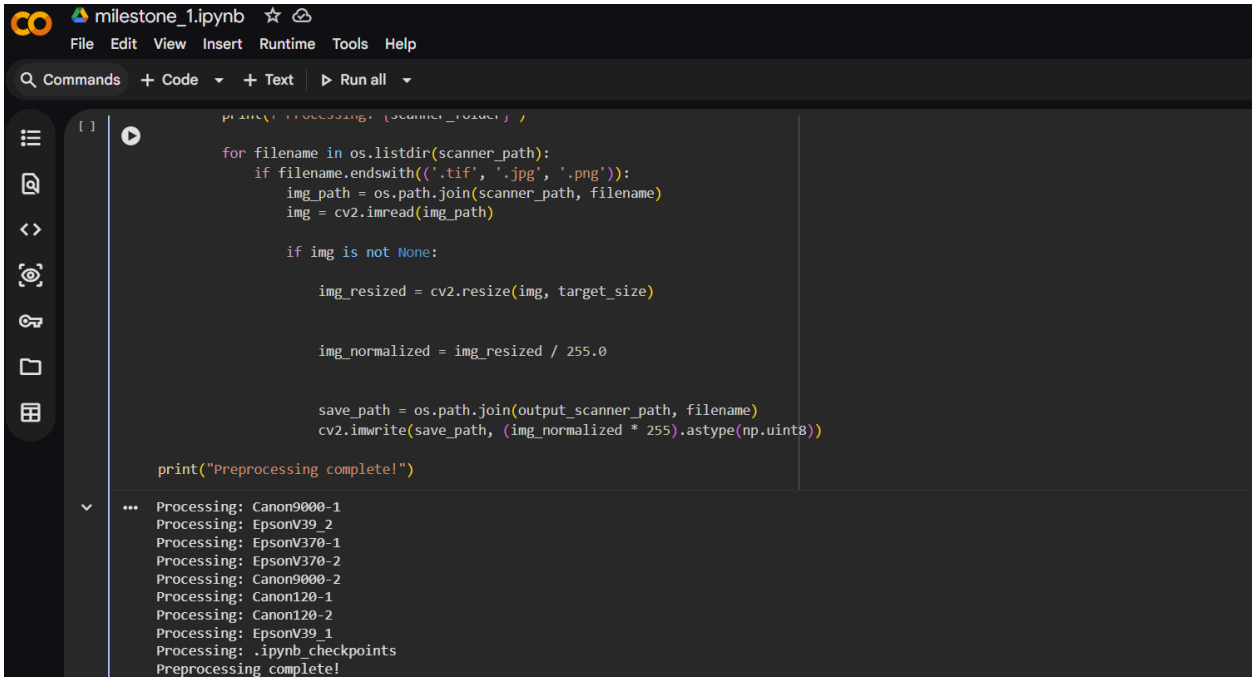
To prepare the data for the machine learning model, the following preprocessing steps were implemented:

- **Resizing:** All images were resized to a fixed dimension for consistency.
- **Normalization:** Pixel values were normalized to structure the dataset for training.

- **Organization:** The dataset has been structured to allow a classifier to distinguish among multiple scanners.

5. Conclusion

Milestone 1 is now complete. The dataset is prepared and ready for Milestone 2, which will involve extracting scanner-specific features such as noise patterns and frequency domain signals (FFT).


milestone_1.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
[ ]
for filename in os.listdir(scanner_path):
if filename.endswith(('.tif', '.jpg', '.png')):
img_path = os.path.join(scanner_path, filename)
img = cv2.imread(img_path)
if img is not None:
img_resized = cv2.resize(img, target_size)
img_normalized = img_resized / 255.0
save_path = os.path.join(output_scanner_path, filename)
cv2.imwrite(save_path, (img_normalized * 255).astype(np.uint8))
print("Preprocessing complete!")
Processing: Canon9000-1
Processing: EpsonV39_2
Processing: EpsonV370-1
Processing: EpsonV370-2
Processing: Canon9000-2
Processing: Canon120-1
Processing: Canon120-2
Processing: EpsonV39_1
Processing: .ipynb_checkpoints
Preprocessing complete!