AI TRACEFINDER

Document Source Identification using Scanner Pattern

Milestone – 2 Report

Infosys Springboard Internship

Submitted by:

Anshul Kumaria

1. Introduction

The primary objective of this milestone is to analyse scanner-specific characteristics using hand-crafted features and to assess the performance of classical machine learning models.

We use a PRNU-based correlation approach to overcome the limitations observed in traditional feature-based models. This step aims to improve scanner identification accuracy by leveraging sensor-level noise patterns.

2. Objectives

The objectives of Milestone-2 are:

- To extract hand-crafted features that capture scanner-specific characteristics.
- To evaluate baseline machine learning models for scanner classification.
- To analyse the limitations of statistical and texture-based features.
- To implement a PRNU-based correlation method for robust scanner identification.

3. Extraction

Several features were extracted from each preprocessed grayscale image to characterise scanner behaviour.

3.1. Intensity-Based

- Mean Intensity: Represents average brightness across the image.
- Standard Deviation: Measures variation in pixel intensities.

3.2. Fast Fourier Transform (FFT) Feature:

The average magnitude of the Fourier transform was computed to capture frequency-domain artefacts introduced by scanner hardware.

3.3. Local Binary Pattern (LBP):

LBP mean and standard deviation were extracted to represent fine- grained texture variations caused by scanner mechanisms.

3.4. Noise Variance Feature

A Gaussian filter was applied to each image, and the residual noise was obtained by subtracting the filtered image. The variance of this noise map was used as a feature to approximate scanner sensor noise.

3.5.PRNU Statistical Features

From the PRNU noise residuals, the following statistical features were extracted:

- PRNU variance
- PRNU energy
- PRNU mean

4. Baseline Machine Learning Models

Using the extracted features, the following classical machine learning models were implemented:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest

The dataset was split into training and testing sets, and model performance was evaluated using classification accuracy.

4.1.Baseline Results

Model	Accuracy
Logistic Regression.	~74%
SVM	~76%
Random Forest	~75%

* Note - The performance of baseline models remained limited due to the loss of spatial sensor noise information.

5. PRNU-Correlation Based Scanner Identification

To address the limitations of feature-based machine learning models, a PRNU-correlation approach was implemented.

5.1.PRNU Residual Extraction

For each image:

- The central region was selected to reduce boundary noise.
- Gaussian filtering was applied to suppress image content.
- The PRNU noise residual was obtained and mean-normalized.

5.2.Scanner-Level Fingerprint Generation

- PRNU residuals from multiple images belonging to the same scanner model were averaged.
- This averaging process reduced random noise and amplified consistent sensor-specific patterns.
- The resulting averaged PRNU served as a unique fingerprint for each scanner model.

5.3.Correlation-Based Matching

- Normalised cross-correlation was computed between test image PRNU residuals and stored scanner fingerprints.
- The scanner model with the highest correlation score was selected as the predicted source.

6. Experimental Results

The PRNU-correlation approach significantly outperformed baseline machine learning models.

Method	Accuracy
Best Baseline Model (SVM)	~76%
PRNU Correlation (Proposed)	**96.97%**

The results clearly demonstrate that correlation-based PRNU matching is more effective for scanner identification than statistical feature-based classifiers.

7. Discussion

The experiments highlight the following observations:

- Hand-crafted features combined with classical ML models provide moderate accuracy.
- Statistical summaries of PRNU are insufficient to capture full sensor characteristics.
- PRNU-correlation preserves spatial noise patterns, leading to significantly higher accuracy.

This confirms that scanner identification is best addressed using sensor-level noise analysis rather than purely statistical learning.

8. Outcome of Milestone-2

By the end of Milestone-2, the following outcomes were achieved:

- Successful extraction of scanner-specific hand-crafted features.
- Implementation and evaluation of baseline machine learning models.
- Identification of limitations in feature-based classification.
- Development of a robust PRNU-correlation approach.
- Achieved high scanner identification accuracy of 96.97%.