

TraceFinder: Forensic Scanner Identification

Introduction

TraceFinder is a project that aims to identify which scanner device produced a particular scanned document. Every scanner leaves behind small, invisible patterns and noise that act like its signature. By studying these patterns in scanned images, it becomes possible to tell which scanner was used. The project uses images from multiple scanners and applies image processing and machine learning techniques to learn these unique scanner characteristics. The final goal is to build a system that can accurately recognize the scanner model based only on the scanned image.

Milestone 1 – Dataset Collection & Preprocessing

1. Introduction

This milestone focuses on understanding the scanned-document dataset, analyzing basic properties, and preparing a consistent preprocessing pipeline. The dataset contains images captured from multiple scanner devices at 150 DPI and 300 DPI.

The goal of Milestone 1 is to clean, normalize, and structure the dataset so that it is ready for feature extraction and model training in later stages.

1. Scanner Models Present in the Dataset

The dataset includes 11 scanner models, each contributing Flatfield, Official, and Wikipedia image samples:

- 1. Canon120-1**
 - 2. Canon120-2**
 - 3. Canon220**
 - 4. Canon9000-1**
 - 5. Canon9000-2**
 - 6. EpsonV39-1**
 - 7. EpsonV39-2**
 - 8. EpsonV370-1**
 - 9. EpsonV370-2**
 - 10. EpsonV550**
 - 11. HP**
-

1. Dataset Description

The dataset is divided into three main components:

1.1 Dataset – 1

Blank pages scanned by each device at 150 and 300 DPI.

These images are crucial for analyzing scanner noise patterns (PRNU-like behaviours).

1.2 Dataset - 1

Real printed-and-scanned documents containing text and graphics.

Available at both 150 and 300 DPI across all scanner models.

1.3 Wikipedia Images

Document pages sourced from Wikipedia, printed, and scanned using the same scanners.

These images add variety and prevent overfitting to official document styles.

All files are stored as TIFF images.

2.Dataset Exploration

Before preprocessing, several image properties were analyzed across Flatfield, Official, and Wikipedia datasets:

- **Resolution (width × height)**
- **Pixel intensity range**
- **Mean and standard deviation**
- **Entropy (image randomness)**
- **Skewness (brightness distribution)**

Key Observations

- **150 DPI images $\approx 1752 \times 1240$ resolution**
- **300 DPI images $\approx 3508 \times 2480$**
- **Flatfields have very low entropy because they contain no content**

- **Official and Wikipedia images show higher entropy, containing text**
- **Pixel values commonly lie between 40 and 255, depending on paper tone and scanner calibration**
- **Several images were originally 16-bit, requiring conversion to 8-bit**

All exploration results were saved in CSV files:

- **flatfield_exploration.csv**
 - **official_exploration.csv**
 - **wikipedia_exploration.csv**
-

2.Dataset Index Creation

Three index files were created to organize the dataset:

- **flatfield_preprocessing.csv**
- **official_preprocessing.csv**
- **wikipedia_preprocessing.csv**

Each contains:

- **path → full file path**
 - **scanner_id → device model**
 - **dpi → 150 or 300**
 - **label_id → numeric label assigned to each scanner model**
-

2.Preprocessing Pipeline

A consistent preprocessing strategy was designed for all image types.

1.4 Grayscale Conversion

All images are converted to 8-bit grayscale, regardless of whether they were originally RGB or 16-bit TIFF.

This simplifies processing and focuses on structural and noise-based features.

1.5 Resizing or Cropping to 512 × 512

To ensure a uniform input size:

- **Official & Wikipedia → resized to 512×512**
- **Flatfields → center-cropped to 512×512 (to preserve scanner noise)**

1.6 Normalization

Pixel values are converted to float and normalized to the range [0, 1].

1.7 No Heavy Denoising

Because scanner identification depends on subtle noise patterns, no aggressive denoising was applied.

2. Visualization and Verification

Several random samples from all categories (150 DPI and 300 DPI) were visualized:

- **Original image (full size)**
- **Preprocessed 512×512 version**

Screenshots of these visualizations were stored in the project for documentation.

Random OFFICIAL 300 DPI ORIGINAL (3508, 2480)



Année Universitaire 2020/2021

ATTESTATION DE SECURITE SOCIALE

Concernant tous les salariés

A faire compléter par votre employeur principal

Je soussigné(e)

agissant en qualité de

Nom de l'employeur

Adresse

Déclare rémunérer l'intéressé(e) ci-dessous désigné :

Nom Prénoms

Numéro d'immatriculation à la sécurité sociale

Ci

Et certifie que l'intéressé(e) effectue, au sein de cette entreprise, au moins 900 heures par an :

☐ OUI ☐ NON Si non, nombre d'heures effectuées sur l'année universitaire :

Sa rémunération principale est soumise (Cochez la case correspondante) :

☐ Au régime général

☐ Au régime fonctionnaire de sécurité sociale (avec pension civile) :

☐ Autre régime spécial de sécurité sociale (précisez lequel) :

Fait à :

Signature et cachet de l'employeur principal :

Le :

La présente attestation n'est valable que pour l'année scolaire en cours.

Random OFFICIAL 300 DPI PREPROCESSED (512, 512)



Année Universitaire 2020/2021

ATTESTATION DE SECURITE SOCIALE

Concernant tous les salariés

A faire compléter par votre employeur principal

Je soussigné(e)

agissant en qualité de

Nom de l'employeur

Adresse

Déclare rémunérer l'intéressé(e) ci-dessous désigné :

Nom Prénoms

Numéro d'immatriculation à la sécurité sociale

Ci

Et certifie que l'intéressé(e) effectue, au sein de cette entreprise, au moins 900 heures par an :

☐ OUI ☐ NON Si non, nombre d'heures effectuées sur l'année universitaire :

Sa rémunération principale est soumise (Cochez la case correspondante) :

☐ Au régime général

☐ Au régime fonctionnaire de sécurité sociale (avec pension civile) :

☐ Autre régime spécial de sécurité sociale (précisez lequel) :

Fait à :

Signature et cachet de l'employeur principal :

Le :

La présente attestation n'est valable que pour l'année scolaire en cours.

Wikipedia 300 DPI ORIGINAL (3508, 2480)



Wikipedia 300 DPI PREPROCESSED (512, 512)

