Deepfake Video Detection

Piyush Dangi | 23EE65R08
Signal Processing and Machine Learning

Under the supervision of

Dr. Rajiv Ranjan Sahay

Department of Electrical Engineering

Outline

1. Introduction
2. Deepfake Detection Methods
3. Photo-Response Non-Uniformity Based Detection Method
4. Datasets
5. Results
6. Future Work

Introduction

- **What are Deepfakes?**
  - Deepfakes are synthetic media, typically videos or images, where a person's likeness is manipulated using AI.

- **Why is deepfake detection important?**
  - Misinformation
  - Security Threats

Deepfake Detection Methods

Traditional Methods based on:

- Image Power Spectrum[1].
- Color Filter Array (CFA)[2].
- Photo-Response Non-Uniformity (PRNU)[3].

Methods based on:

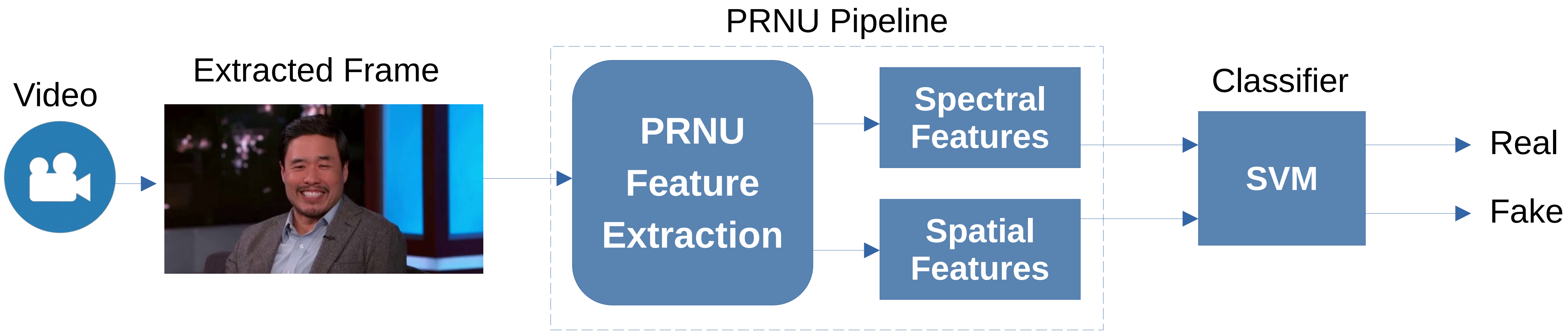
- Biometric Feature
- Deep learning

[1] Durall, R., Keuper, M. and Keuper, J., 2020. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7890-7899).

[2] Gallagher, A.C. and Chen, T., 2008, June. Image authentication by detecting traces of demosaicing. In 2008 IEEE computer society conference on computer vision and pattern recognition workshops (pp. 1-8). IEEE.

[3] Chen, M., Fridrich, J., Goljan, M. and Lukás, J., 2008. Determining image origin and integrity using sensor noise. IEEE Transactions on information forensics and security, 3(1), pp.74-90.

PRNU Based Deepfake Detection

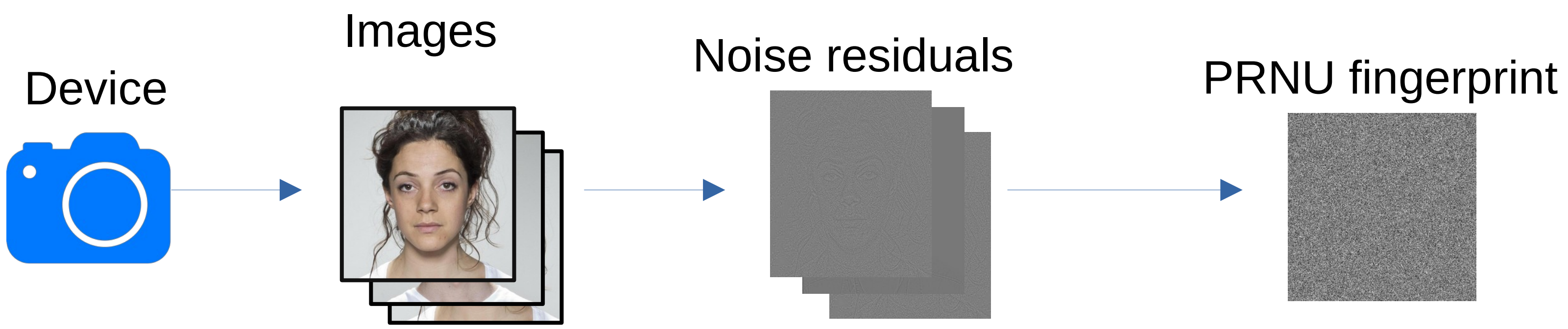


Overview of the pipeline structure of the PRNU based deepfake detection

PRNU: Photo-Response Non-Uniformity
SVM: Support Vector Machine

Photo-Response Non-uniformity

- Due to manufacturing imperfections, camera sensor elements exhibit small deviations from expected behavior, forming a stable noise-like pattern known as Photo-Response Non-Uniformity (PRNU).



The device PRNU Pattern is estimated by a large number of noise residuals.

PRNU Extraction

- The extraction of the PRNU noise residual from an image is performed by applying Fridrich's approach[3].
- For each image $\mathbf{I}_k$ the noise residual $\mathbf{W}_k$ is estimated as described in Eq. (1),

$$\mathbf{W}_k = \mathbf{I}_k - F(\mathbf{I}_k) \qquad (1)$$

The estimation of the PRNU $\mathbf{K}$:

$$\mathbf{K} = \frac{\sum_{k=1}^{N} \mathbf{W}_k \mathbf{I}_k}{\sum_{k=1}^{N} (\mathbf{I}_k)^2} \qquad (2)$$

Where
F : Denoising function
$\mathbf{I}_k$: Image
$\mathbf{W}_k$: Noise residual

N : number of images obtained by the camera.

[3] Chen, M., Fridrich, J., Goljan, M. and Lukás, J., 2008. Determining image origin and integrity using sensor noise. IEEE Transactions on information forensics and security, 3(1), pp.74-90.

PRNU Splitting

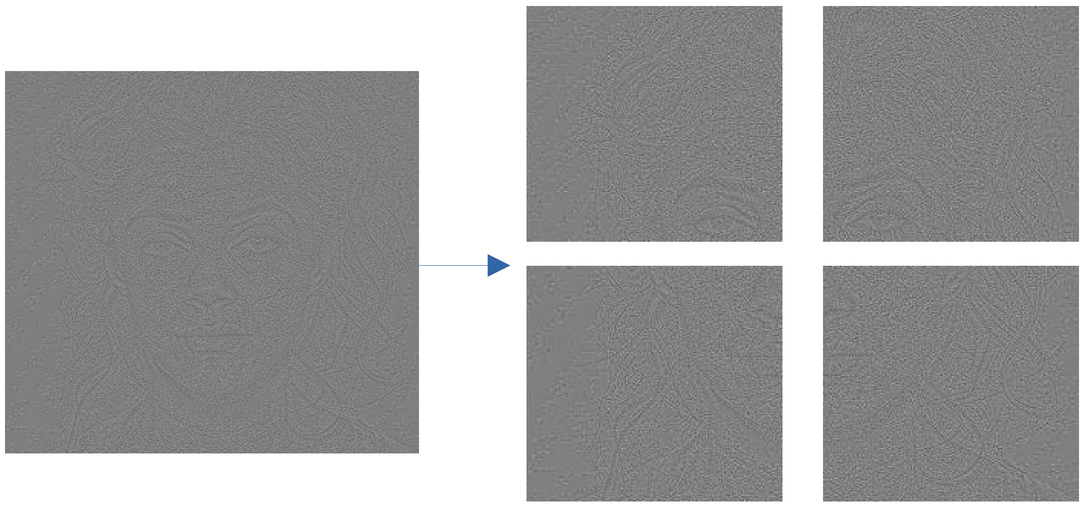


Figure 1: Example for splitting the PRNU into $C = 4$ Cells (2×2) of equal size.

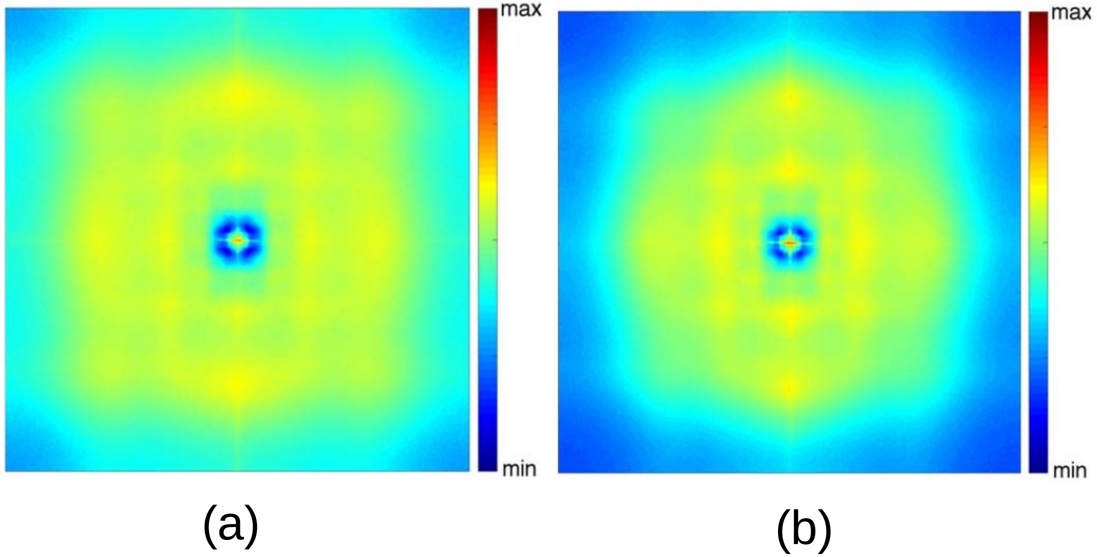


Figure 2: DFT magnitude spectra of the PRNUs extracted from (a) real (b) fake face images.

Spatial Features

- P_{en} Energy of the PRNU values:

$$P_{en} = \sum_{i=1}^{N} |x_i|^2 \tag{3}$$

- P_{skew} Skewness of PRNU values:

$$P_{skew} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)^{\frac{3}{2}}} \tag{4}$$

Where:
- N is the number of PRNU values.
- x_i represents each PRNU value.
- μ is the mean of the PRNU values.

Spatial Features

- P_{kurt} Kurtosis of PRNU values:

$$P_{kurt} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)^2} - 3 \tag{5}$$

- P_{varH} Variance of values in PRNU histogram:

$$P_{varH} = \frac{1}{B}\sum_{n=1}^{B}\left(H_p(n) - \bar{H}_p\right)^2 \tag{6}$$

Where:

- N is the number of pixels.
- x_i represents each pixel value.
- μ is the mean of the pixel values.
- B is number of bins in PRNU cell's histogram.
- $H_p(n)$ PRNU cell's histogram.
- $\bar{H}_p$ represents the mean frequency of the histogram bins.

Spatial Features

- P_{maxH} Position of maximum value in PRNU histogram:

$$P_{maxH} = \underset{n=1,\dots,B}{\operatorname{argmin}} H_p(n) \quad (7)$$

Where :

- B is number of bins in PRNU cell's histogram
- $H_p(n)$ PRNU cell's histogram

- For the histogram features P_{varH} and P_{maxH} , we created a histogram H of PRNU values of a cell,

Similarly, spectral features can be calculated from the DFT magnitude spectra of the PRNU cell,

PRNU Features

Spatial Feature:-

P_{en} : Energy of PRNU values

P_{var} : Variance of PRNU values

P_{skew} : Skewness of PRNU values

P_{kurt} : Kurtosis of PRNU values

P_{varH} : Variance of values in PRNU histogram

P_{maxH} : Position of maximum value in PRNU histogram

Spectral Feature:-

D_{en} : Energy of DFT values

D_{var} : Variance of DFT values

D_{skew} : Skewness of DFT values

D_{kurt} : Kurtosis of DFT values

D_{varH} : Variance of values in DFT histogram

D_{maxH} : Position of maximum value in DFT histogram

Feature Aggregation

- For every feature, we compute the mean and variance over all cells of one image.

Example:-

Mean of feature P_{en} :

$$\bar{P}_{en} = \frac{1}{C}\sum_{c=1}^{C} P_{en}(c) \tag{8}$$

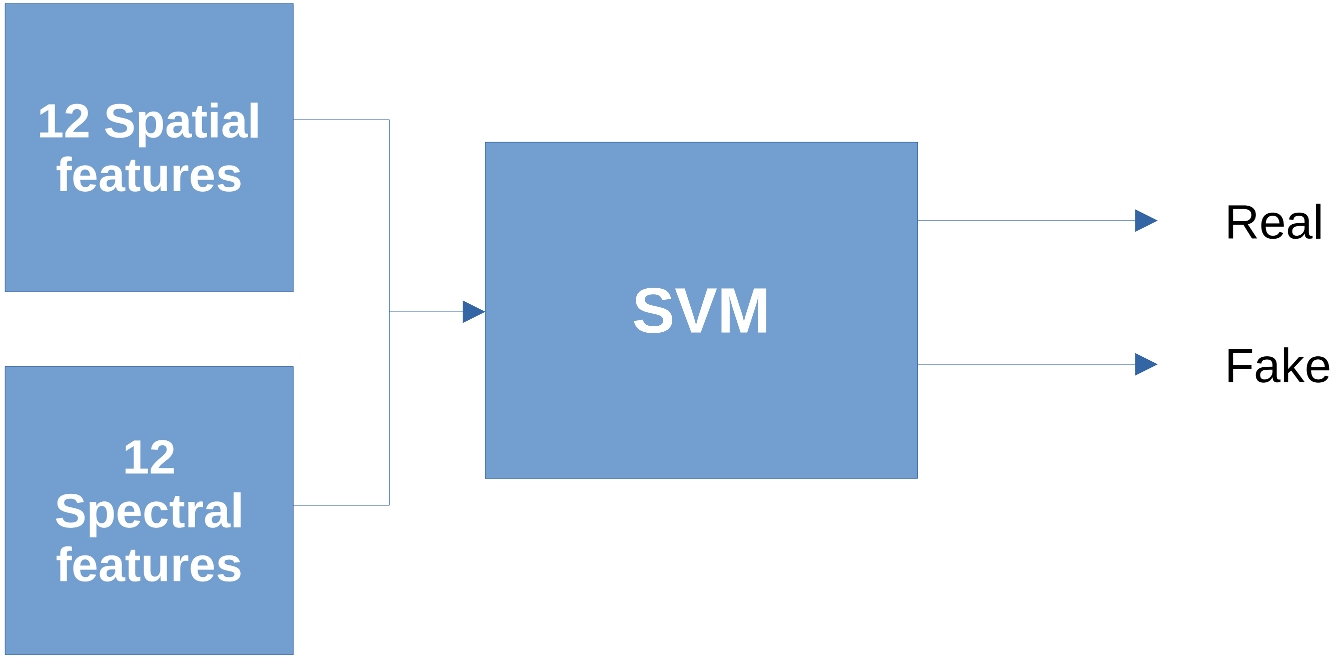
Variance of feature P_{en} :

$$P_{enVar} = \frac{1}{C}\sum_{c=1}^{C} \left(P_{en}(c) - \bar{P}_{en}\right)^2 \tag{9}$$

Where:
- C is number of cells.
- $P_{en}(c)$ represents energy of PRNU values of a cell.

- So, this results in a total of 24 features per image.

Classification



Datasets

Dataset	Pristine / Forged	Frame size	Year
FaceForensics++ [4]	1,000 / 4,000	480p, 720p, 1080p	2019
Celeb-DF [5]	590 / 5,639	various	2020
DeeperForensics-1.0 [6]	50,000 / 10,000	1080p	2020
DFDC [7]	19,154 / 100,000	240p - 2160p	2019

[4] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M., 2019. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1-11).

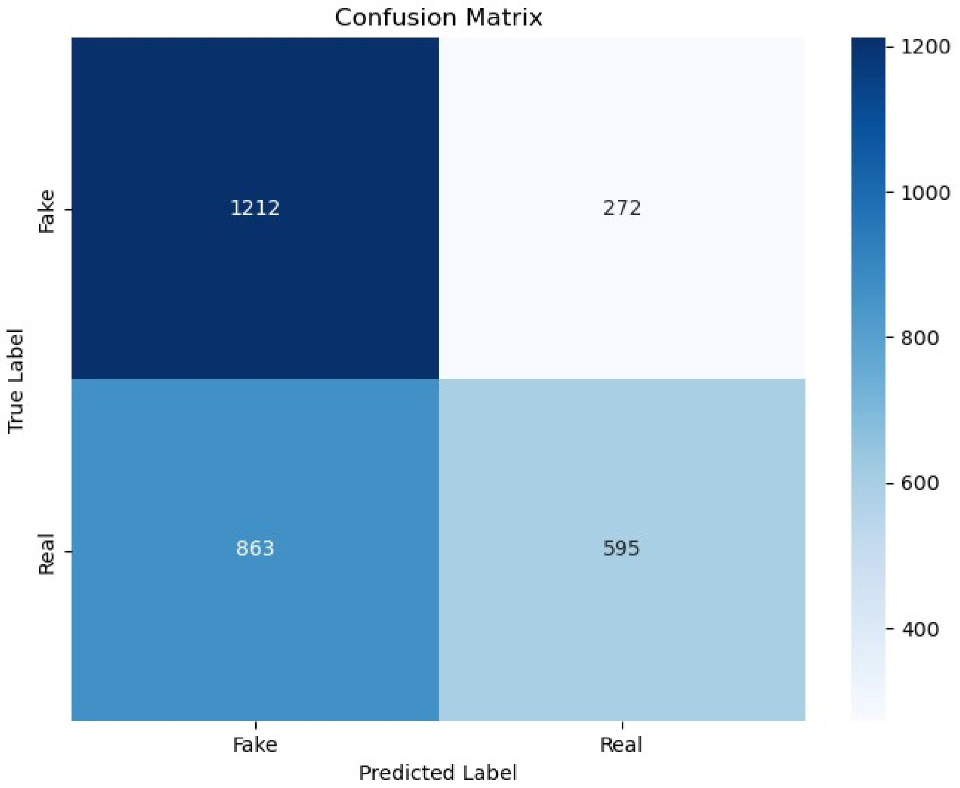
[5] Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S., 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3207-3216).

[6] Jiang, L., Li, R., Wu, W., Qian, C. and Loy, C.C., 2020. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2889-2898).

[7]Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M. and Ferrer, C.C., 2020. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397.

Results

- Accuracy $\approx 61.3\,\%$
- Error rate $\approx 38.5\,\%$



Future Work

Literature Review for Deepfake Detection methods → **To implement Fully Unsupervised based deepfake video detection** → **To develop light weight network for deepfake detection**

July 2024 – August 2024 | September 2024 – November 2024 | December 2024 – March 2025

Thank you